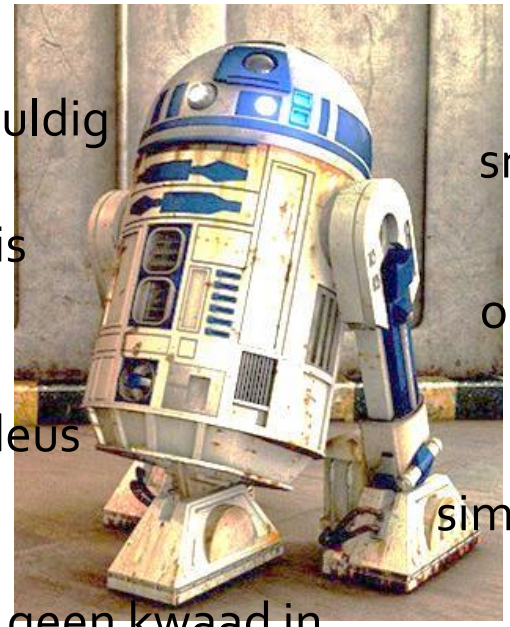
Concept features and lexical heterogeneity in dialects

Karlien Franco

supervisors:
Dirk Geeraerts
Dirk Speelman
Roeland van Hout

ONSCHULDIG
‘innocent’



onschuldig

kuis

zebedeus

daar zit geen kwaad in

snulletje

onnozel

simpel

18 different words
(Lim. & Brab.)

?

BANGERIK
‘coward’

bange

bangerd

angstige

bange floets

schouwe

bange pezerik

held op sokken

angstpiemel

bangboks

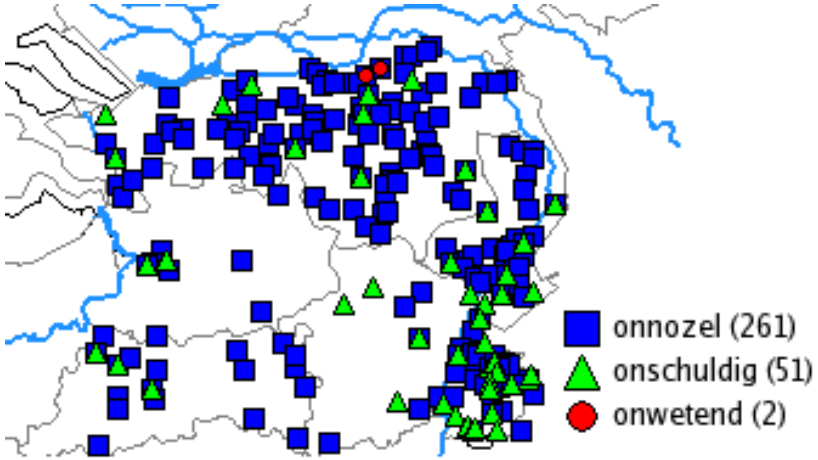
schrikkepee

angstschijter

bang schijthuis

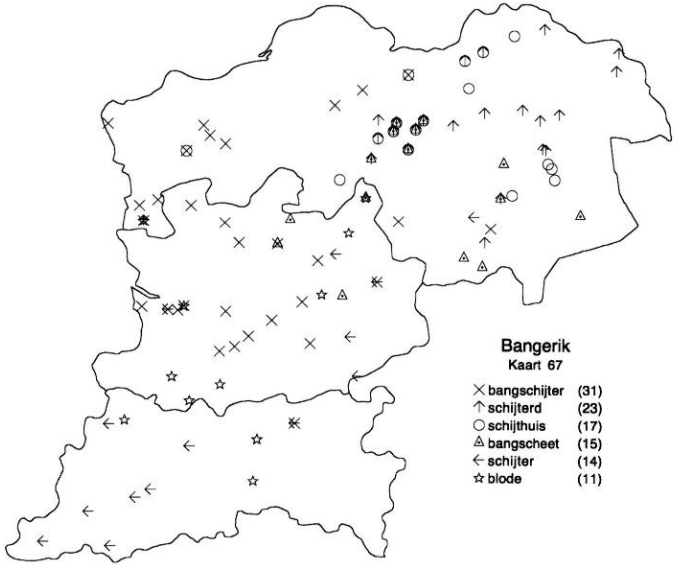
100 different words
(Lim. & Brab.)

ONSCHULDIG
'innocent'



two variants that occur nearly everywhere

BANGERIK
'coward'



small geographical areas

?

pilot studies

- concept characteristics influence the amount of lexical dialect variation

- more lexical geographical variability for concepts that
  - are prone to negative affect
  - have a low degree of onomasiological salience
  - are vague

(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)

negative affect (Limburg)

WELL BUILT WOMAN (GROF GEBOUWDE VROUW)	
machochel	mokkel
schommel	bai (fr.)
m	
m	
kapitein	dikke madam
mangel	dikke prij
machochel	flink wijf
schommel	fors vrouwmens
molenpaard	bammel
…	…

HEAD (HOOFD)
hoofd
kop

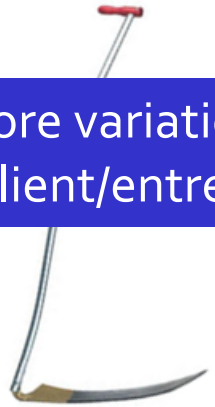
significantly more variation for concepts that are prone to negative affect

onomasiological salience

"various categories may have various degrees of entrenchment"

(Geeraerts, Grondelaers & Speelman 1999: 8)

e.g. CABLE TIES CUTTER vs. SCYTHE vs. SCISSORS



significantly more variation for concepts
that are less salient/entrenched/familiar

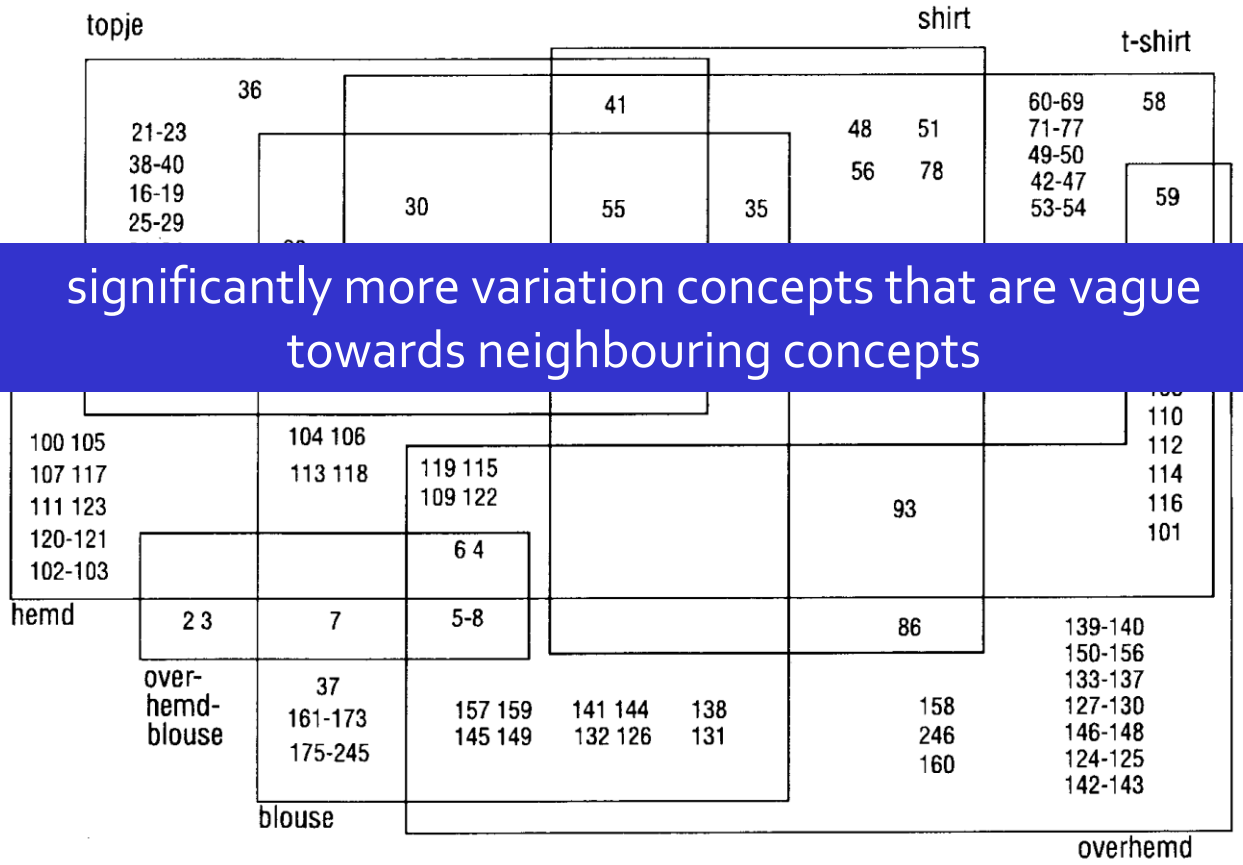
onomasiological salience

lack of salience (Limburg)

LITTLE DENTS BETWEEN THE KNUCKLES (KNOKKELKUILTJES)	
boelenhandjes	kuiltjes
deukjes	kussens
dompels	kwabbel
kinkdraaier	lokje
knobbels	plooien
knokkelkuiltjes	putjes
knokkels	vetkuiltjes
knookjes	vingerkotjes
kotjes	vouwen
kreukeling	vouwtjes

HEAD (HOOFD)
hoofd
kop

onomasiological vagueness



significantly more variation concepts that are vague towards neighbouring concepts

non-discreteness in the lexical field of shirt-like garments
(Geeraerts, Grondelaers & Bakema 1994: 140)

vagueness (Limburg)

MODEST (INGETOGEN)	
bedaard	niet opvallend
bedeesd	onopvallend
bescheiden	op zijn eigen
charmant	ruhig (du.)
deftig	rustig
eenvoudig	serieus
fatsoenlijk	simpel
gemtlich (du.)	stemmig
gewoon	stil
ingetogen	teruggetrokken
kalm	zoet
modest	

PEACEFUL, QUIET (KALM, BEDAARD)
bedaard
evenwichtig
gemoedelijk
gemtlich (du.)
kaduuk
kalm
koest
ruhig (du.)
rustig
stil
traag
zoet

vagueness (Limburg)

TUESDAY (DINSDAG)
dinsdag

WEDNESDAY (WOENSDAG)
woensdag
asgoensdag
goensdag
mittwoch (du.)

research questions

why do some concepts show
more lexical geographical variation than others?

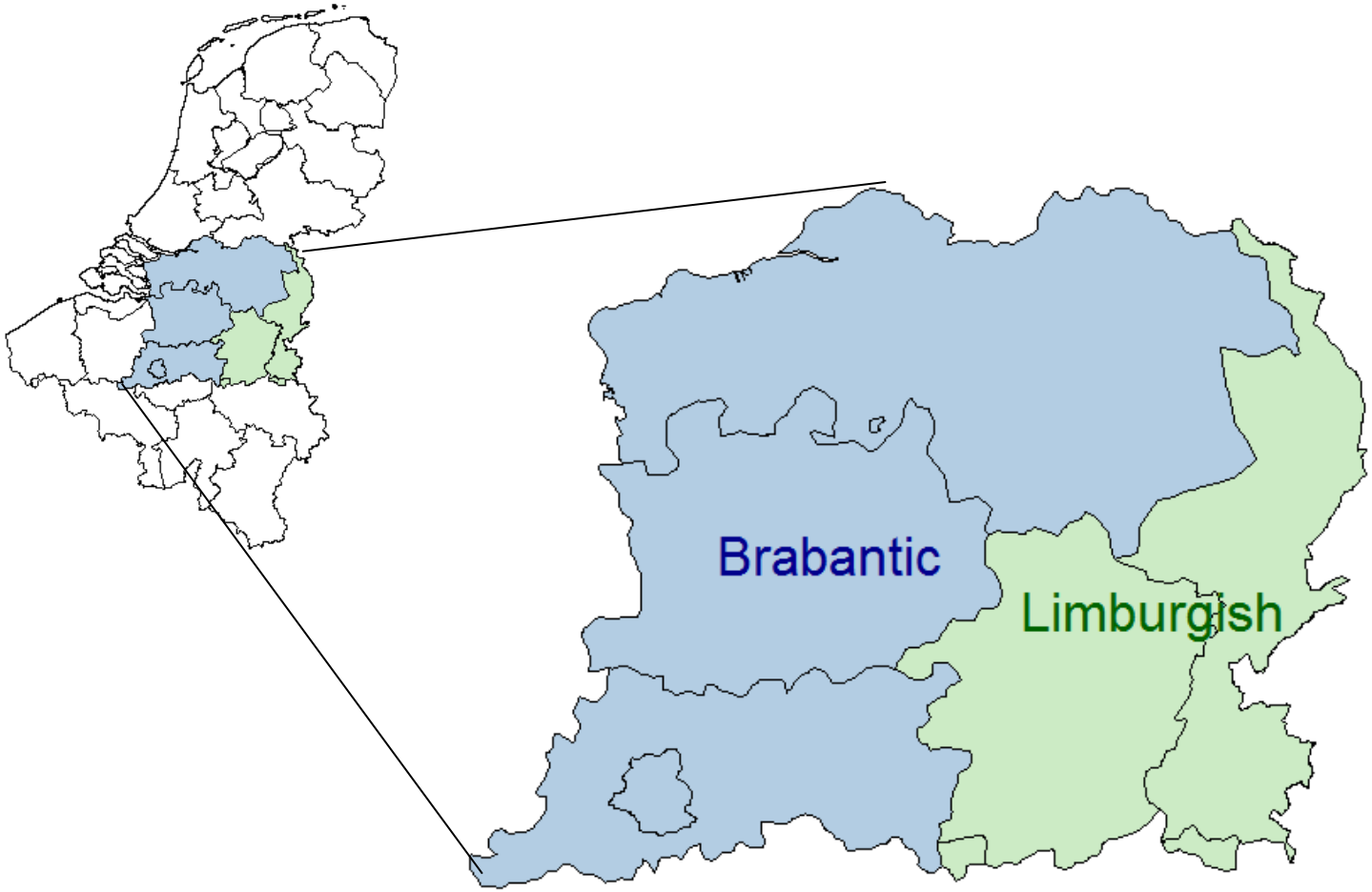
- confirm that the influence of concept-related features is stable in
  - other semantic fields
  - other dialect areas
  - other language areas
  - other types of data
- determine which other features may influence lexical geographical dialect variation

data

- databases of two (three in ch. 6) onomasiological dialect dictionaries:
  - WBD: Woordenboek van de Brabantse dialecten
  - WLD: Woordenboek van de Limburgse dialecten
  - see a.o. Kruijsen 1996 for the history of these dictionary projects
  - case-study 4: WVD (Woordenboek van de Vlaamse dialecten) & DBÖ (database of Bavarian dialects in Austria)

the dialects of Dutch

the dialects of Dutch



subsetting the data

- thematically: part 3 - general vocabulary
  - 14 chapters (WLD & WBD)
  - 1 chapter = 1 semantic field
  - one or more semantic field(s) per case-study

semantic fields (WLD)

PART 3: General vocabulary

1: Man as an individual (De mens als individu)

- The human body (Het menselijk lichaam)
- Physical activity and health (Beweging en gezondheid.)
- Clothing and grooming (Kleding en lichamelijke verzorging)
- Personality and feelings (Karakter en gevoelens)

2: Domestic life (Het huiselijk leven)

- The house (De woning)
- Family and sexuality (Familie en seksualiteit)
- Food and drink (Eten en drinken)

3: Community life (Het gemeenschapsleven)

- Society, school and education (Maatschappelijk gedrag, school en onderwijs)
- Celebration and entertainment (Feest en Vermaak)
- Church and religion (Kerk en geloof)

4: The world versus man (De wereld tgo. de mens)

- Fauna: birds (Fauna: vogels)
- Fauna: other animals (Fauna: overige dieren)
- Flora (Flora)
- The physical and abstract world (De stoffelijke en abstracte wereld)

semantic fields (WLD)

PART 3: General vocabulary

1: Man as an individual (De mens als individu)

- The human body (Het menselijk lichaam)
- Physical activity and health (Beweging en gezondheid.)
- Clothing and grooming (Kleding en lichamelijke verzorging)
- Personality and feelings (Karakter en gevoelens)

2: Domestic life (Het huiselijk leven)

- The house (De woning)
- Family and sexuality (Familie en seksualiteit)
- Food and drink (Eten en drinken)

3: Community life (Het gemeenschapsleven)

- Society, school and education (Maatschappelijk gedrag, school en onderwijs)
- Celebration and entertainment (Feest en Vermaak)
- Church and religion (Kerk en geloof)

4: The world versus man (De wereld tgo. de mens)

- Fauna: birds (Fauna: vogels)
- Fauna: other animals (Fauna: overige dieren)
- Flora (Flora)
- The physical and abstract world (De stoffelijke en abstracte wereld)

subsetting the data

- thematically: part 3 - general vocabulary
  - 14 chapters (WLD & WBD)
  - 1 chapter = 1 semantic field
  - one or more semantic field(s) per case-study

- practically: only data collected by NCDN through questionnaires
  - only concepts > 50 places
  - only places > 50 concepts

    → systematicity

from questionnaire …

KEUKEN-EN KOOKGEREI 2.

1. Hoe noemt u in uw dialekt:
   Het metalen voorwerp met hengsel en tuit dat dient om water in te koken?

   a. het moderne voorwerp (zie ill.) .....

2. b. het oudere voorwerp, dikwijls met ronde bodem, dat boven een haardvuur werd gehangen of in de opening van een kachel werd gezet (zie ill.)
   .....

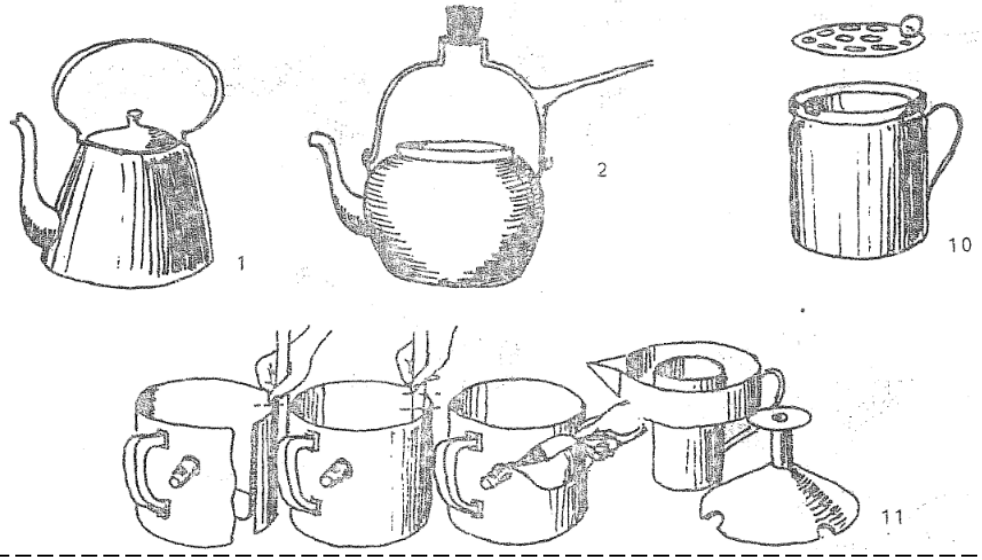
3. De pot met deksel, tuit en oor waarin koffie wordt gezet?
   .....

4. De zak waardoor het water over de gemalen koffie wordt gegoten?
   .....

5. De buikige pot met tuit en oor waarin enkel gezet?
   .....

6. Wanneer werd er thee gedronken? Als dagelijk als geneeskrachtige drank?

WOORDENBOEK VAN DE VLAAMSE DIALEKTEN : bijlage bij WVD 43 : huisraad
(Gelieve de illustraties niet terug te sturen)
Het nummer van de illustratie komt overeen met het nummer van de vraag.



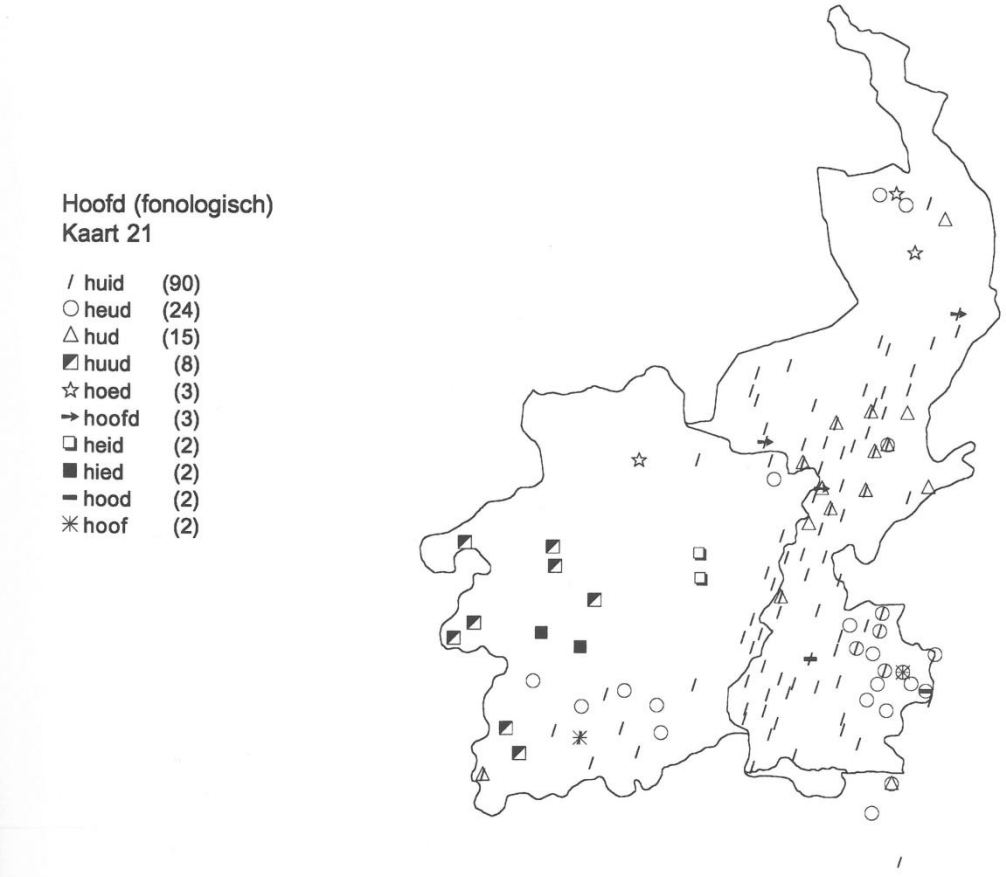
… to dataset …

concept	**variant**	**question**	**location**	**…**
damesmantel ‘coat for women’	caban (fr.)	damesmantel, inventarisatie uitdrukkingen	Tervuren	…
overjas ‘overcoat’	frak	een jas die men over het colbert heen draagt	Leopoldsburg	…
…	…		…	…
vrolijk ‘cheerful’	spass (du.) haan	een opgeruimde, lichte, blijde stemming […]	Simpelveld	…
vrolijk ‘cheerful’	opgewekt	een opgeruimde, lichte, blijde stemming […]	Venlo	…
…	…		…	…

… to measurements at the level of the concept

concept	lexical geographical variation	predictor 1: affect sensitivity	predictor 2: vagueness	…
achterdochtig 'suspicious'	5	sensitive	2.275	
achterhoofd 'back of the head'	21	neutral	4.977	…
…	…	…		…
speelplaats 'playground'	3	neutral	2.341	…
speels 'light-hearted'	9	sensitive	3.561	…
…	…	…		…
…	…	…		…

NB: phonological variation



four case studies

1. systematization of and extensions on the pilot studies

   → is the influence of concept features stable in other semantic fields and dialect areas?

2. de-stratification

   → is the influence of concept features stable if we control for the geographical signal in the data?

3. excusing my French/Latin/German

   → how does the cultural-historical background of a language user influence lexical dialect variation?

4. let's talk about plants, baby

   → what is the influence of the everyday environment of a language user on lexical dialect variation?

1.

concept features influence lexical geographical variation

systematization of and extensions on the pilot studies

replication of pilot studies

SYSTEMATIZATION

→ effect of concept characteristics in other fields than the human body and in other dialect areas

EXTENSION

→ other influential factors?

- individual vs. community (e.g. Pickl 2013)
- concrete vs. abstract concepts

data: design

	concrete	**abstract**
man as an individual	the human body (4.390)	personality and feelings (2.347)
domestic life	the house (4.345)	family and sexuality (3.359)
community life	celebration and entertainment (3.772)	society, school and education (3.260)

data: design

	concrete	**abstract**
man as an individual	the human body (4.390)	personality and feelings (2.347)
domestic life	the house (4.345)	family and sexuality (3.359)
community life	celebration and entertainment (3.772)	society, school and education (3.260)

(mean concreteness: Brysbaert et al. 2014)

concept-related predictors

1. LACK OF SALIENCE
   - proportion of missing places
     → ambiguous
   - proportion of multi-word expressions (MWE)
   - proportion of hapax legomena
   - prevalence (Keuleers et al. 2015)
     - word-level
     - missing data
2. VAGUENESS
   - number of types also used for other concepts (GS10, SG08)
3. AFFECT
   - manual, but relatively stable
   - mean valence (Moors et al. 2013), but missing data

components of lexical dialect variation

- lexical diversity

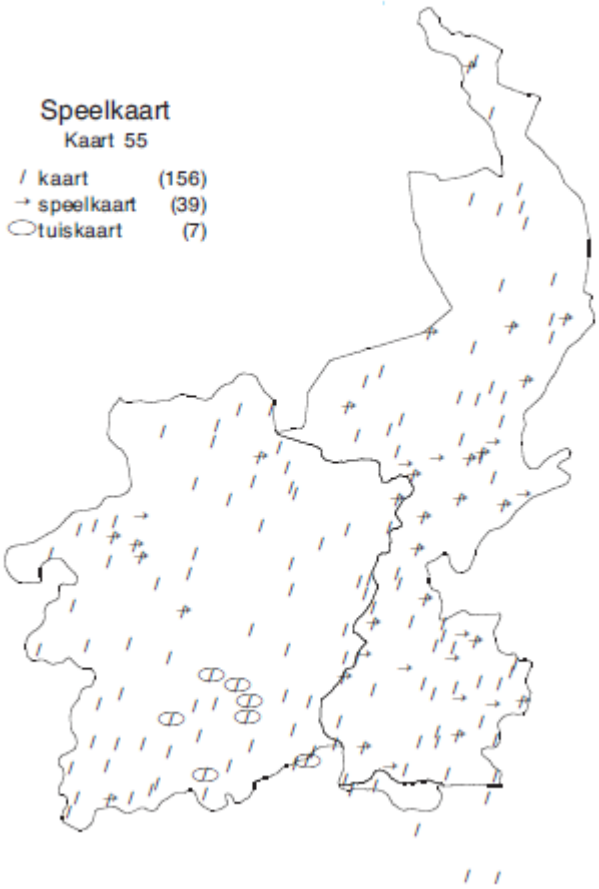
  → some concepts have more different dialectal variants than others

- geographical fragmentation
  - dialect data is geographical in nature
  - geographical scatter of variants can range from very homogeneous to very heterogeneous

→ log(lexical diversity * geographical fragmentation)

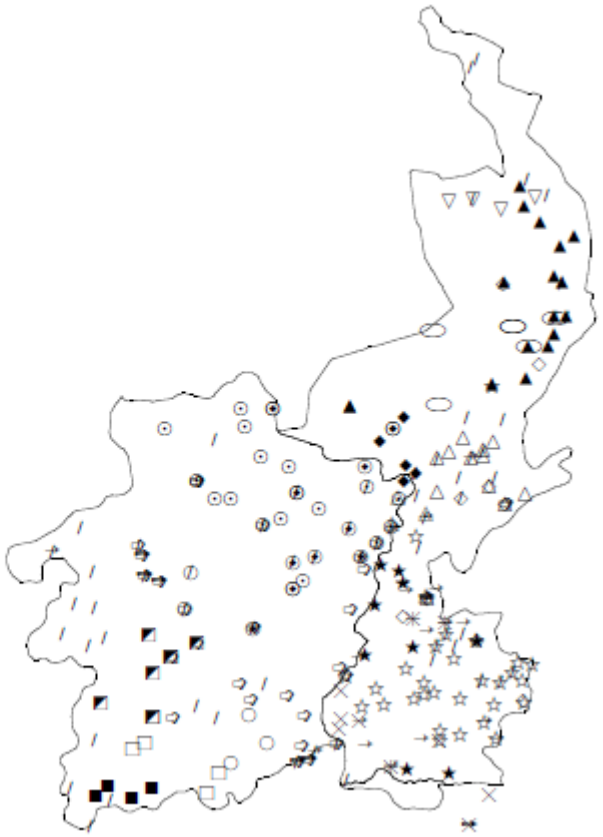
(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)

homogeneous vs. heterogeneous



Verstoppertje spelen
Kaart 17

/ verstoppertje spelen	(63)
→ versteken	(22)
⇢ versteken spelen	(7)
⇨ verstekertje spelen	(27)
◇ verbergen	(9)
△ bergmannetje spelen	(12)
◬ bergemannetje spelen	(4)
▲ bergemuisje spelen	(17)
▽ piepmuis spelen	(4)
○ piepbergen	(6)
⊙ piepjebergen	(18)
⦿ piepjeverbergen	(11)
◆ stoppiepje verbergen	(5)
☆ koekverbergen	(29)
★ koekversteken	(9)
× koekepiep spelen	(10)
+ piepekoek spelen	(4)
⬭ stokverbergen	(6)
□ lonken	(6)
◩ lonkertje spelen	(6)
■ lonkkat spelen	(4)



method

- linear regression
- adjusted R^2 = 0.6756
- formula (significant effects only):

lexical heterogeneity ~

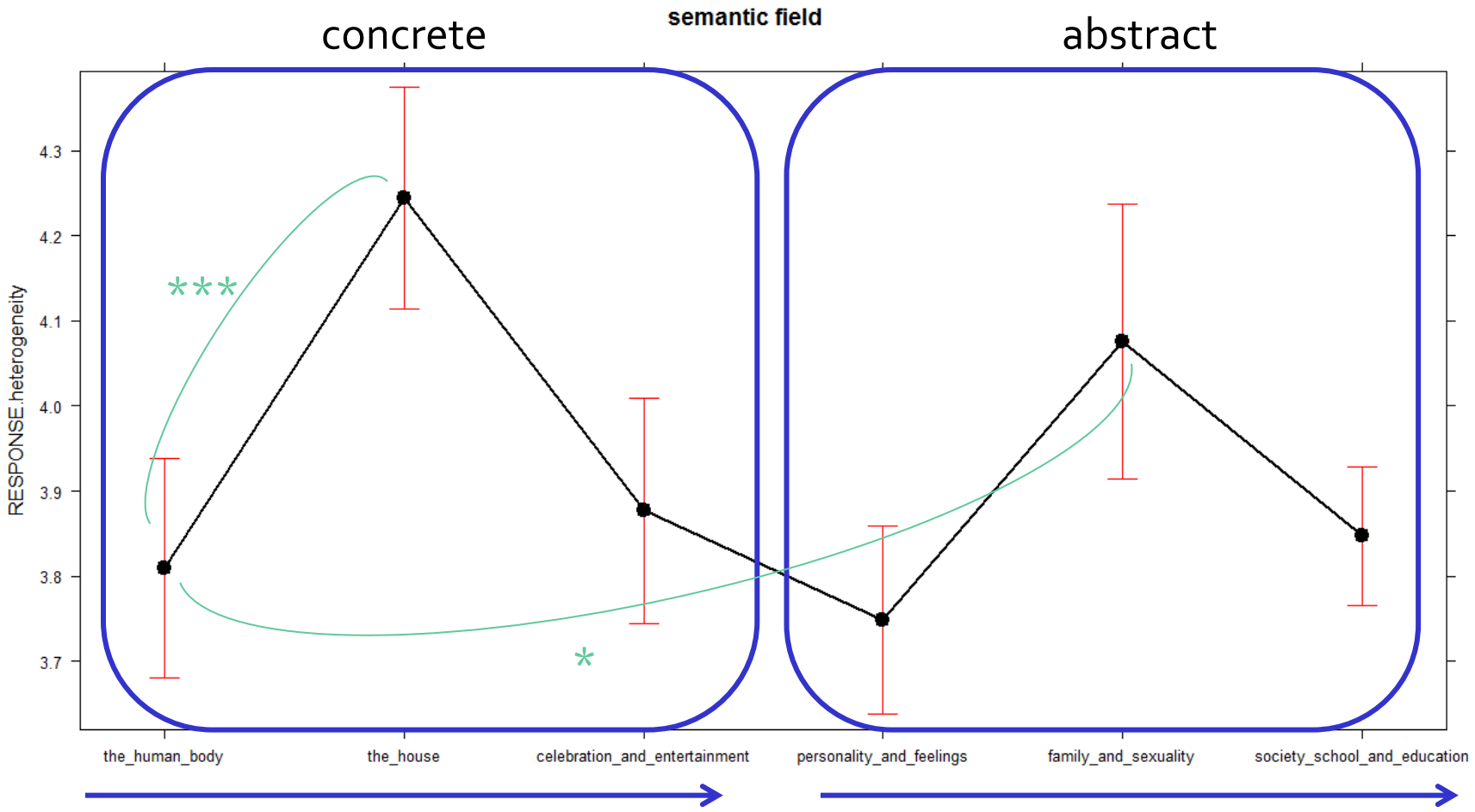
semantic field +

lack of salience (prop. of MWE's + prop. of hapaxes)

vagueness +

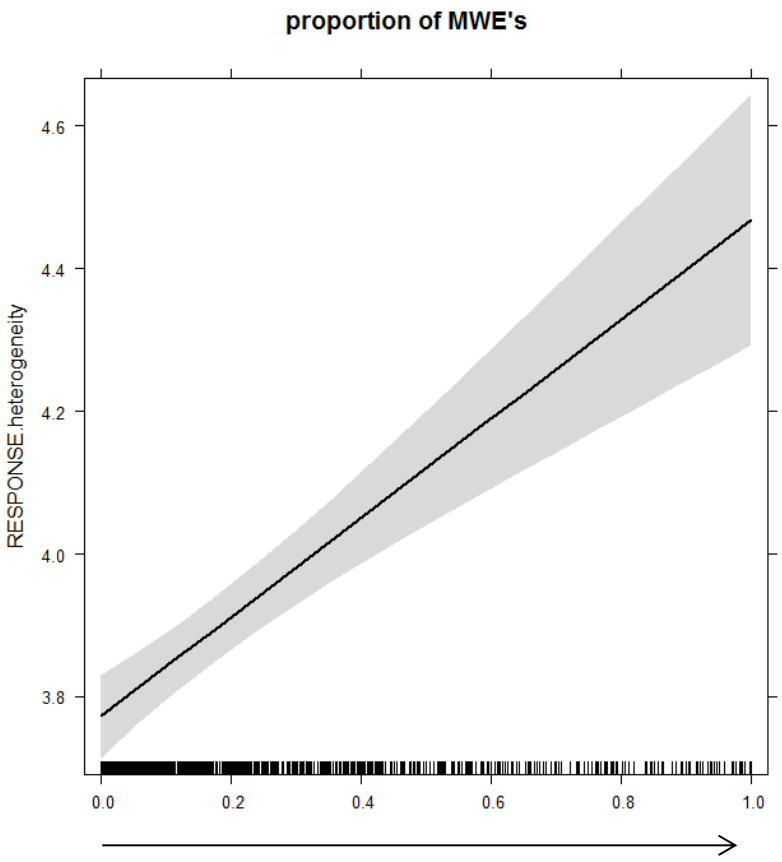
affect (manual coding)

results
semantic field

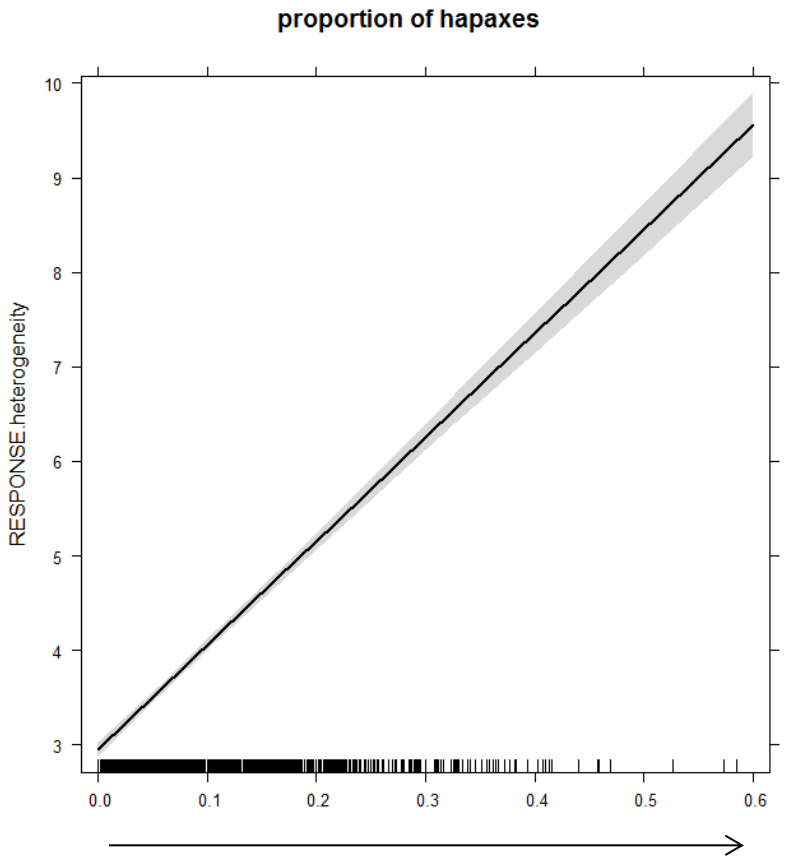


local > society-related > universal?

results
lack of salience

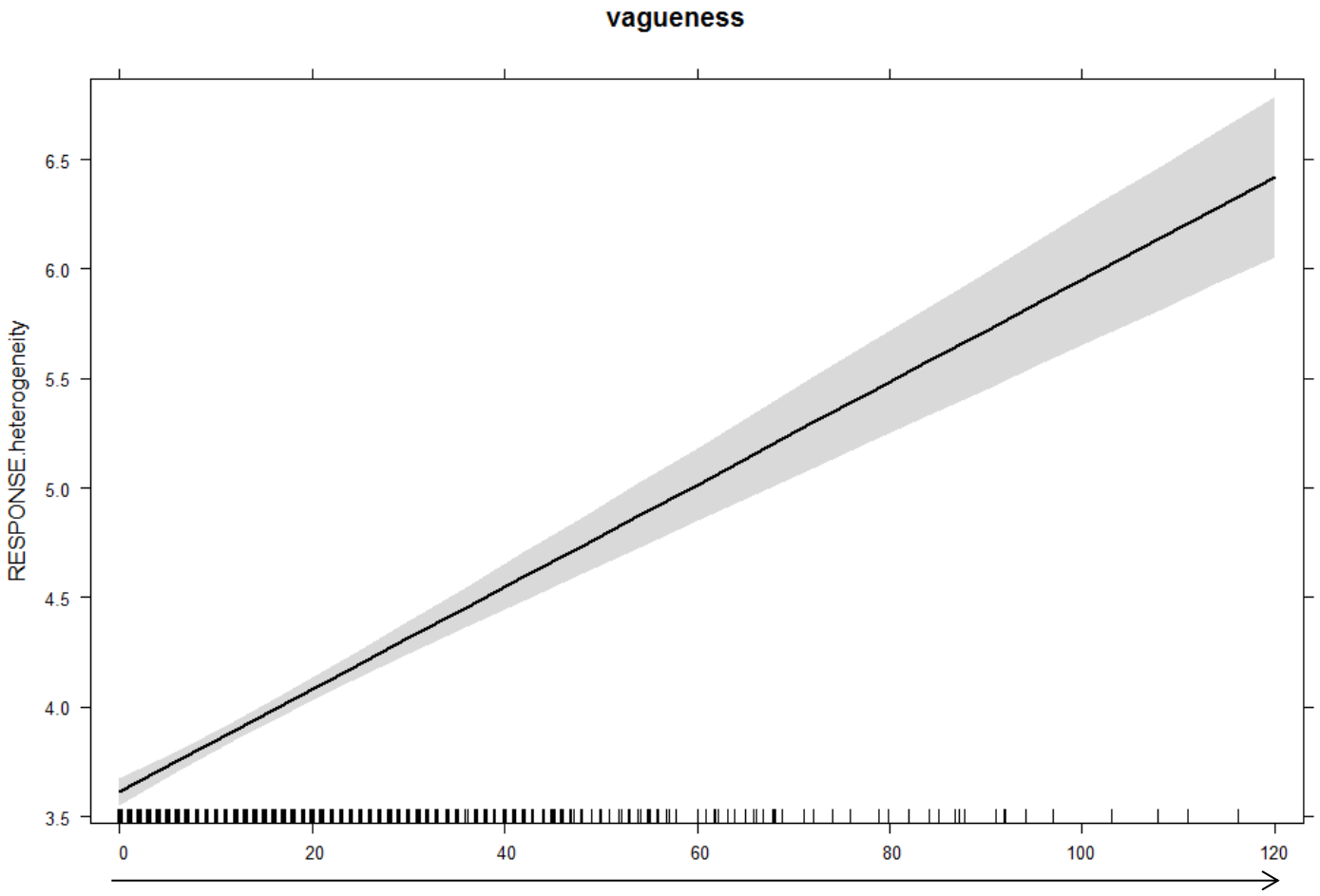


lack of salience

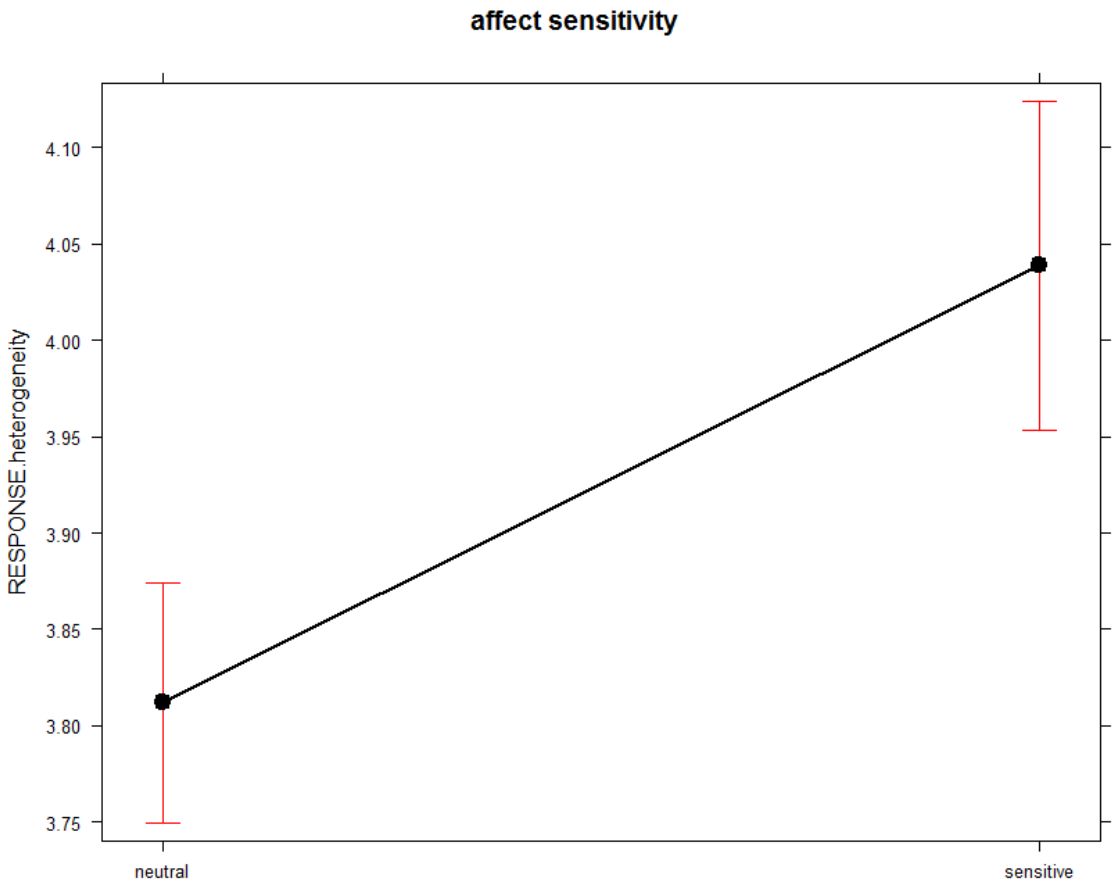


lack of salience

results
vagueness



results
affect sensitivity



discussion

SYSTEMATIZATION

lack of salience, vagueness and affect also lexical dialect variation in other fields than the human body

EXTENSION

- no clear effect of concreteness

  → on the concept-level?

- local > society-related > universal

to do

- affect
- other dialect area: WBD

2.

de-stratifying the data

measuring the influence of concept features

on the lexical component

RESIDUALIZED

research questions

do concept characteristics also influence variation in the lexicon-at-large?

two possible methodologies:

- data stratified along a different dimension than geography
- control for the geographical signal in dialect data

research questions

do concept characteristics also influence variation in the lexicon-at-large?

two possible methodologies:

- data stratified along a different dimension than geography
- control for the geographical signal in dialect data

methodology

1. linear regression model:

   lexical diversity ~ geographical fragmentation

   adj. R^2 = 0.4611

   correlation residuals & lexical diversity = 0.310 (spearman)

2. residuals as response variable in second model with concept characteristics as predictors

   → are the results still stable?

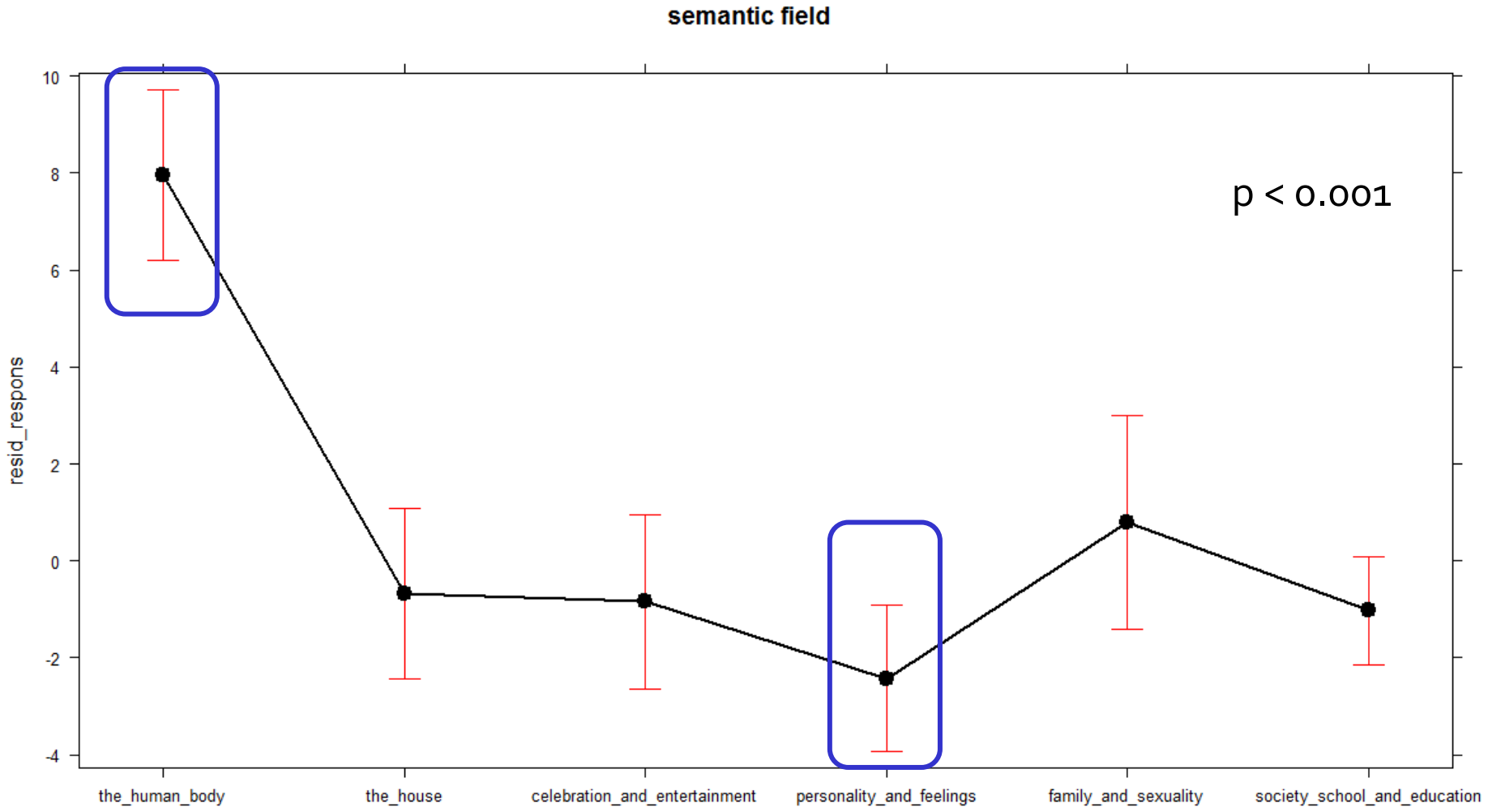
results

- model formula identical

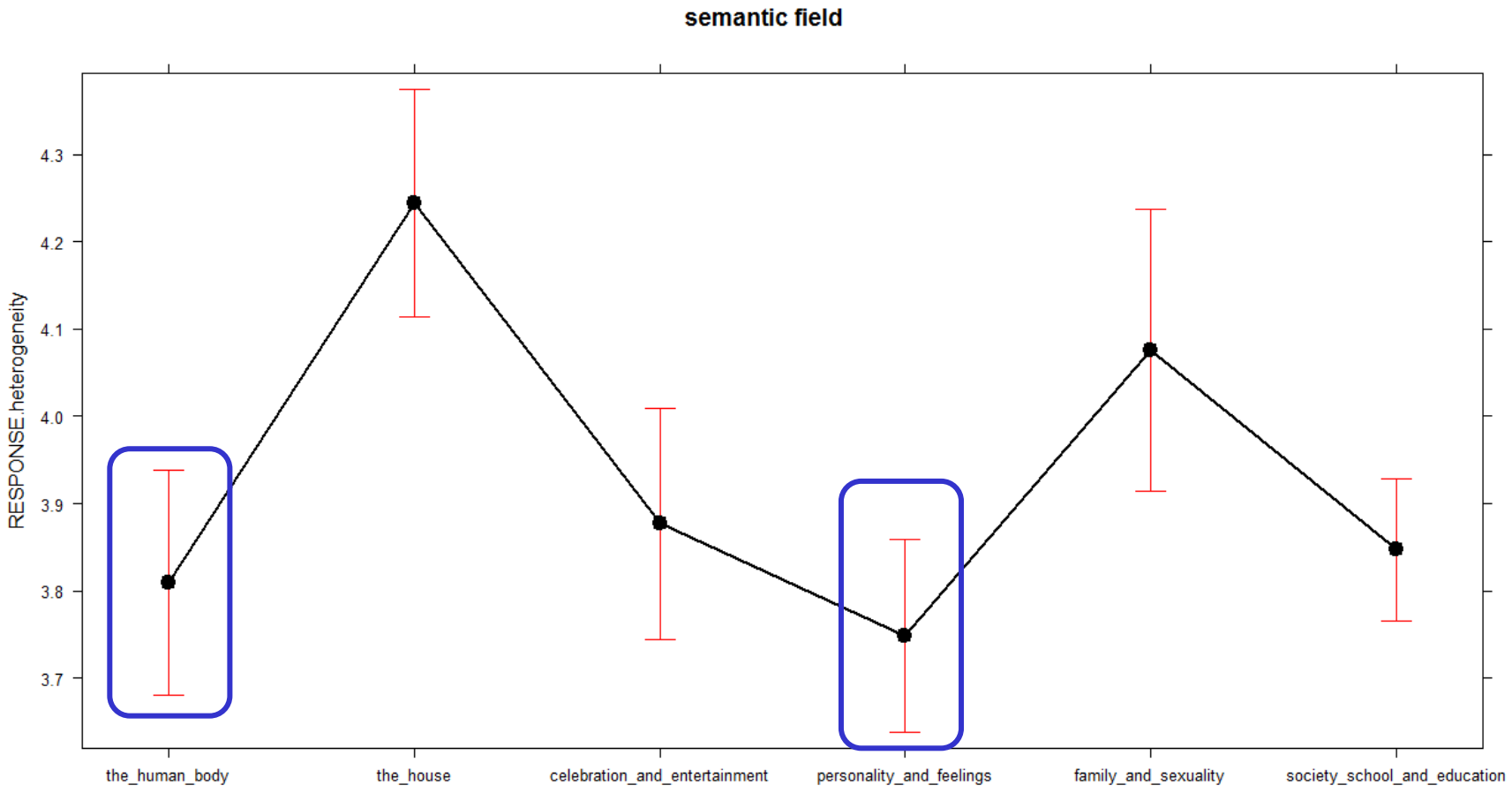
- concept features all have significant effect
  - more variation for less salient concepts
  - more variation for vaguer concepts
  - more variation for concepts prone to affect

- adj. R^2 much lower (0.2292)

results


semantic field
resid_respons
10
8
6
4
2
0
-2
-4
p < 0.001
the_human_body
the_house
celebration_and_entertainment
personality_and_feelings
family_and_sexuality
society_school_and_education

vs. results case-study 1



discussion

- preliminary results indicate that concept features also influence the lexicon-at-large

  → further research

- clear differences between semantic fields

  → some fields more prone to purely lexical variation

3.

excusing my French / Latin / German

modelling variation in the use of loanwords

in dialectal varieties

there is structure in naming strategies

- names for birds reflect how well-known a bird is
- similar patterns occur for names of clothes
- plant names are often based on the shape or color of the plant
- useful plants (i.e. edible plants or plants with medicinal applications) show less lexical variation (cf. infra)

→ naming strategies show how language users structure their daily environment

(Swanenberg 2000, Geeraerts, Grondelaers & Bakema 1994, Brok 1993)

borrowing as a naming strategy

- necessary and luxury loans

  cheerleader vs. freak (*zonderling*)

- the success of a loanword differs per semantic field

Latin:	a.o.	christianity	e.g. *evangelie*, *kardinaal*, *klooster*
		military	e.g. *defensie*, *pijl*
French	a.o.	ME courts	e.g. *baldakijn*, *buffet*, *kasteel*
		administration	e.g. *parket*, *parlement*
		clothing	e.g. *mannequin*, *jupon*, *bretel*

- diachronic differences

(Van der Sijs 1996, Zenner, Speelman & Geeraerts 2012)

geographical differences in loanword usage

- more intense language contact with French in Flanders than in the Netherlands apparent from the higher number of French loans in Spoken Belgian Dutch

  vb. *camion*, *kravat*, *gazet*

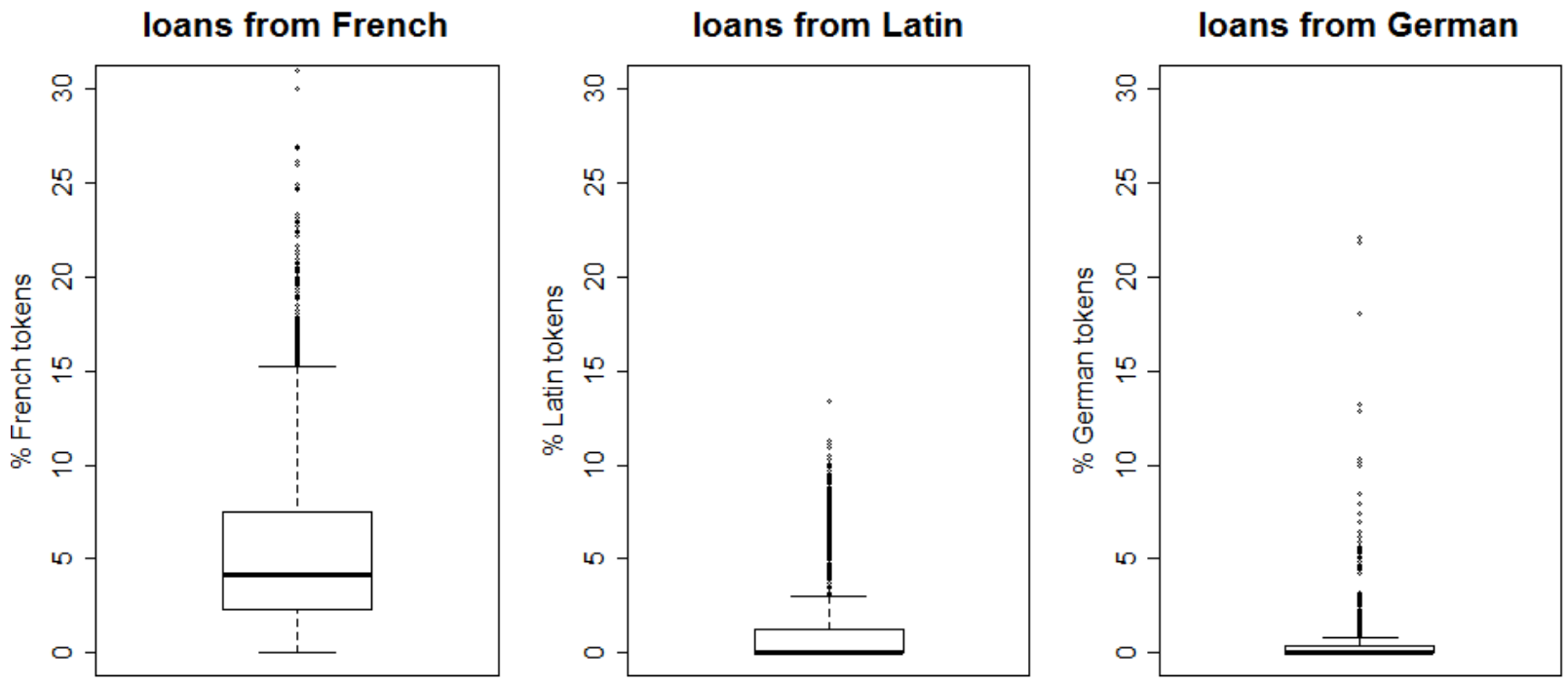
  N.B. purism

- more language contact near language borders

  but state border can evolve into a dialect border

(Weijnen & Van Coetsem 1957, Giesbers 2008, Van der Sijs 1996)

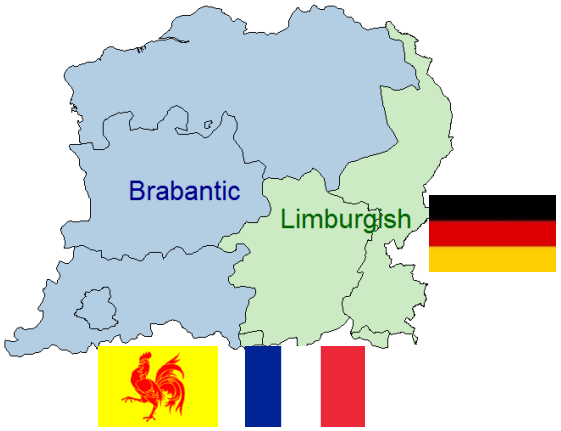
can we find structure in the usage of loanwords?



geographical structure? semantic structure?

we expect…

- geographical patterns



  - French: Flanders > Netherlands
  - German: border effect
  - Latin: no effect

- differences between semantic fields
  - more French for clothing terms and (mostly in Flanders) for concepts relating to society and education
  - more Latin for concepts concerning church & religion

in practice

concept	variant	location	…
damesmantel ‘coat for women’	caban (fr.)	Tervuren	…
overjas ‘overcoat’	frak	Leopoldsburg	…
…	…	…	…
vrolijk ‘cheerful’	spass (du.) haan	Simpelveld	…
vrolijk ‘cheerful’	opgewekt	Venlo	…
…	…	…	…
heilige hostie ‘sacred host’	hostie (lat.)	Bocholt	…
heilige hostie ‘sacred host’	Ons Lieve Heer	Neerpelt	…
…	…	…	

in practice

concept	variant	location	…
damesmantel ‘coat for women’	caban **(fr.)**	Tervuren	…
overjas ‘overcoat’	frak	Leopoldsburg	…
…	…	…	…
vrolijk ‘cheerful’	spass **(du.)** haan	Simpelveld	…
vrolijk ‘cheerful’	opgewekt	Venlo	…
…	…	…	…
heilige hostie ‘sacred host’	hostie **(lat.)**	Bocholt	…
heilige hostie ‘sacred host’	Ons Lieve Heer	Neerpelt	…
…	…	…	

data distribution

543 659 words (tokens)
43 828 different words (types)
2 338 concepts
637 locations
221 368 Brabantic tokens
322 291 Limburgish tokens
29 458 French tokens
10 171 Latin tokens
2 635 German tokens

analyze the proportion of French/Latin/German variants per location

e.g. largest proportion of French occurs in Vorsen (over 30% of all tokens)

- 'combinaison' (ONDERJURK) vs. 'onderrok' & 'onderkleed'
- 'bijou' (JUWEEL) vs. 'juweel' & 'edelsteen'
- 'pardessus' (OVERJAS) vs. 'overjas'

Generalized Additive Modelling (GAM)

- extension of GLMs, which allows for more complex relationships between predictors and response (wiggliness)
- one model per source language (French, Latin, German)
- basic model:

proportion of loanwords per location ~

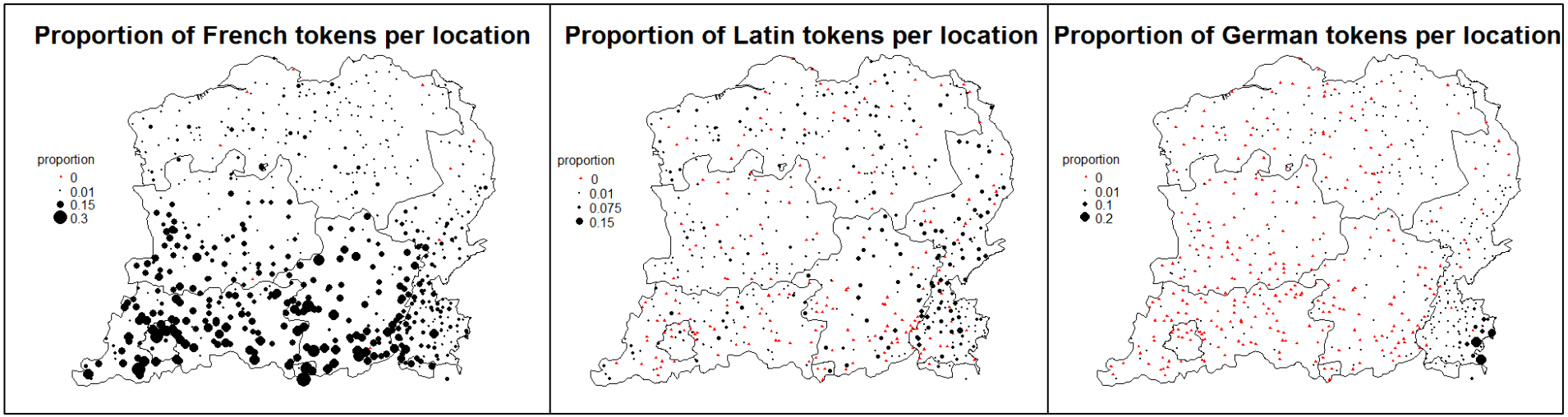
semantic field +

smooth term for lon*lat by semantic field +

random intercept for location (NS for Latin)

(Crawley 2007, Faraway 2006, Wood 2006, Wieling 2012, Zuur et al. 2009)

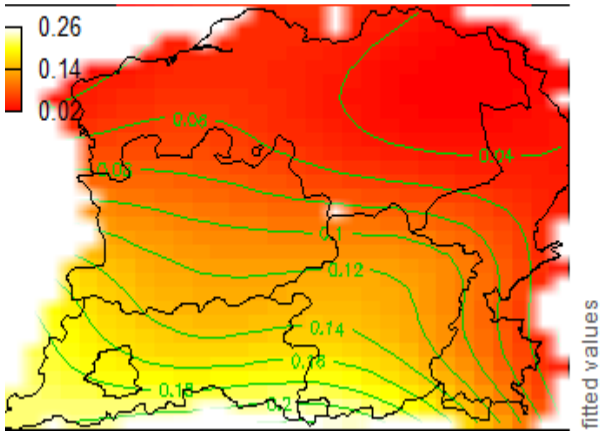
the general picture



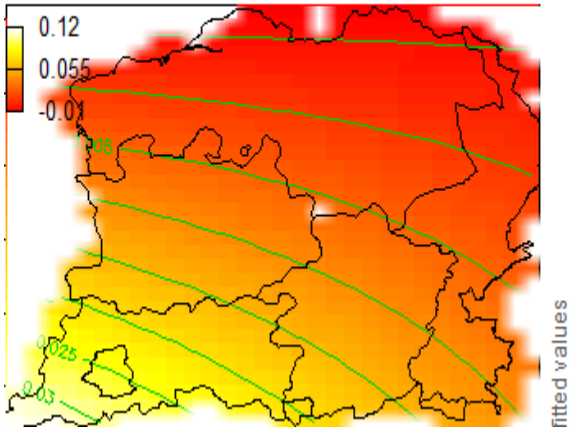
deviance explained: 89.6%

semantic patterns: French

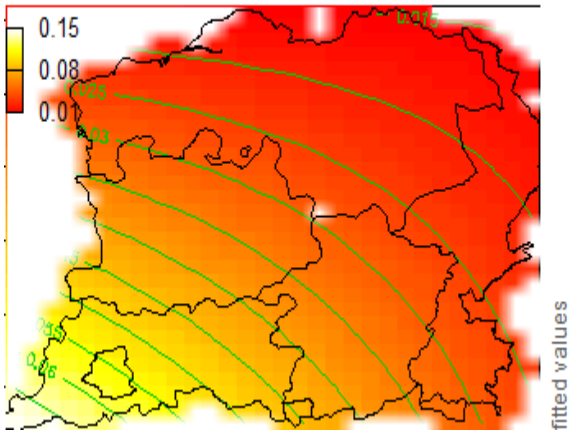
clothing



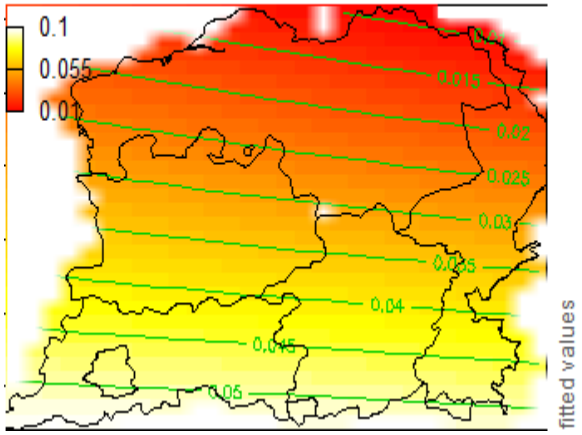
personality & feelings



church & religion



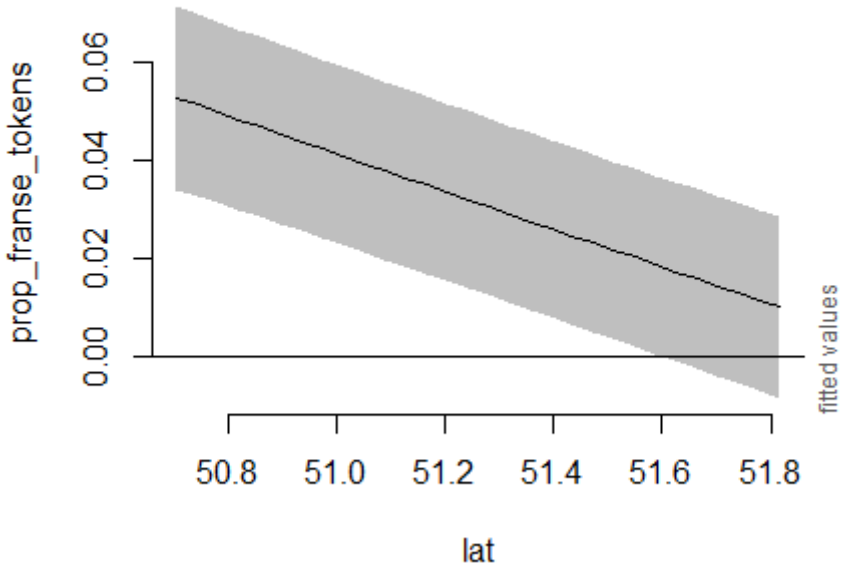
society, school & education



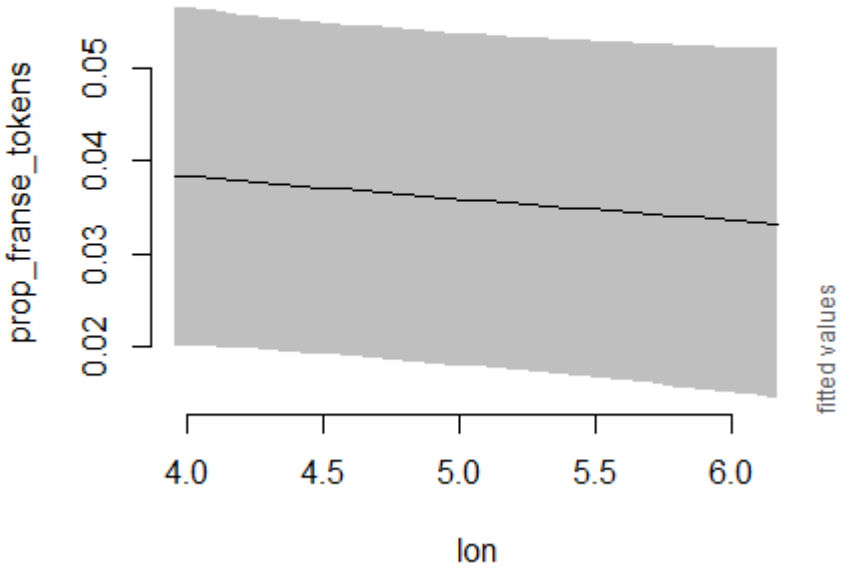
deviance explained: 89.6%

geographical patterns: French

south-north



west-east

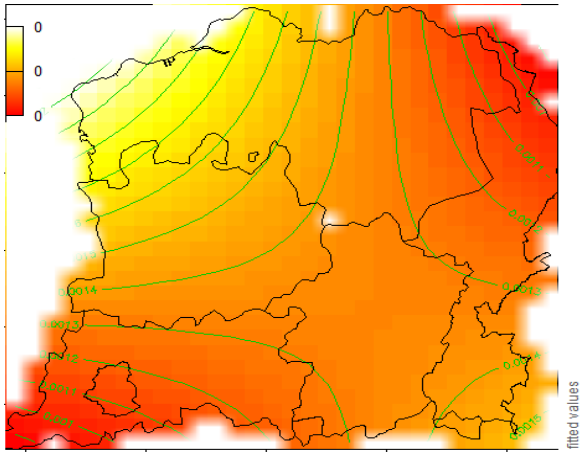


semantic patterns: Latin

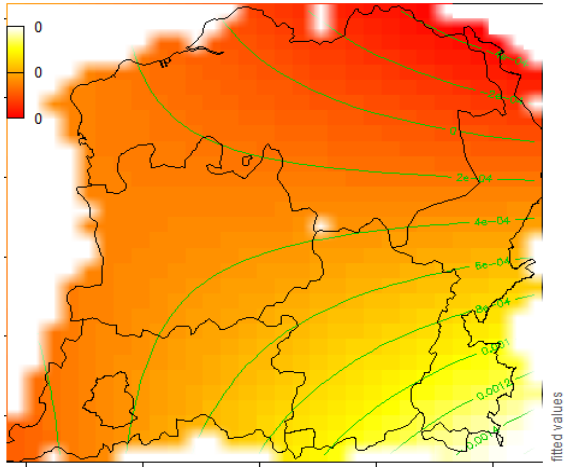
deviance explained : 91.8%

88% without geography

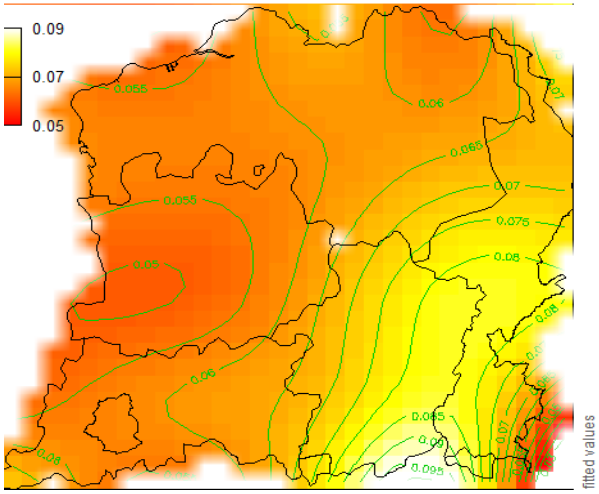
clothing



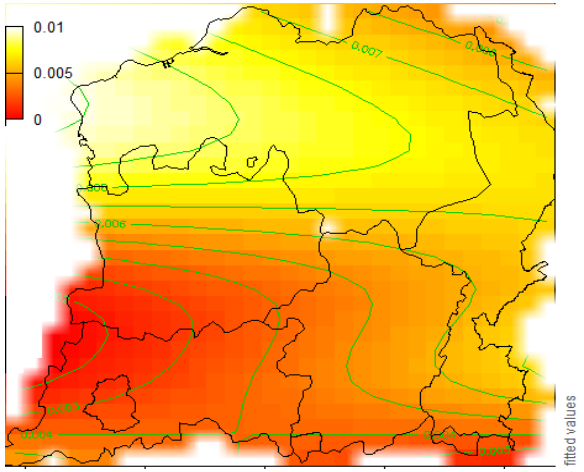
personality & feelings



church & religion



society, school & education

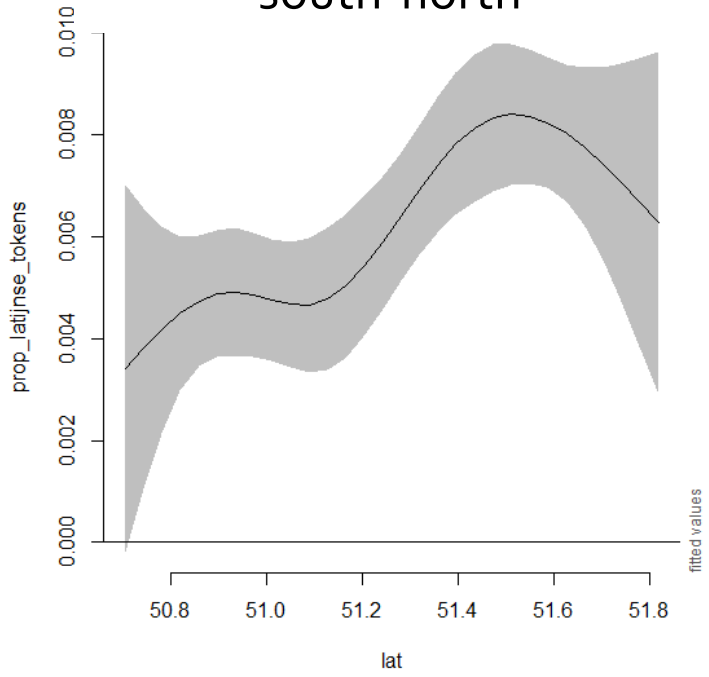


geographical patterns: Latin

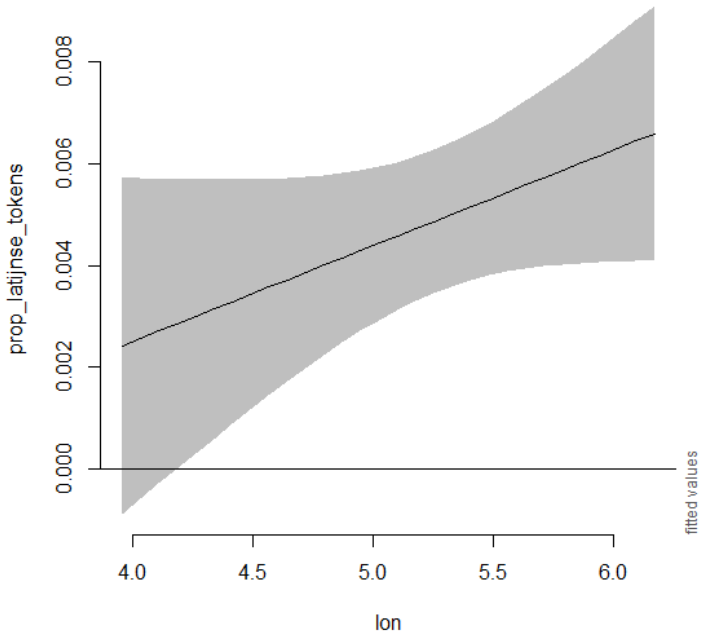
deviance explained : 91.8%

88% without geography

south-north



west-east

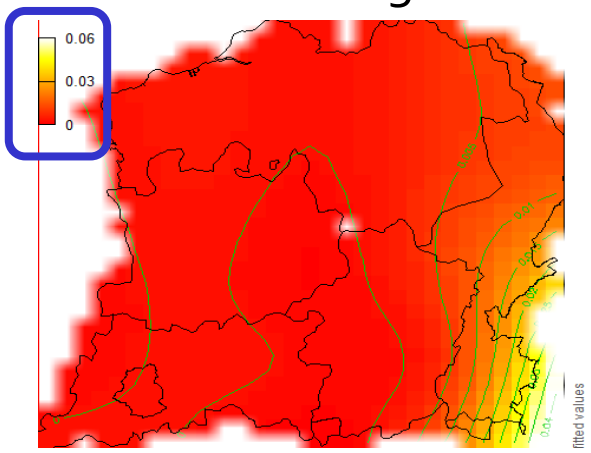


deviance explained : 90.4%

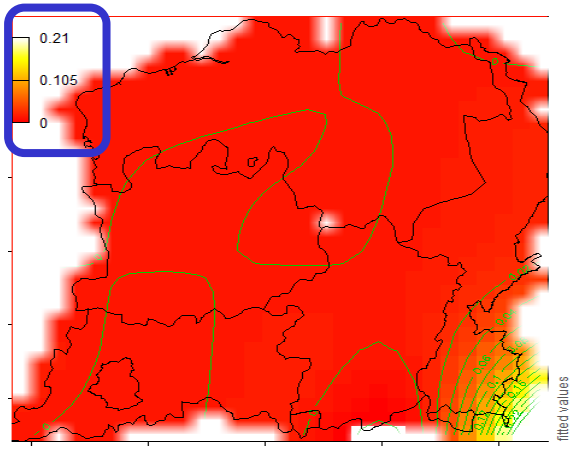
semantic patterns: German

model struggles with general infrequency of German

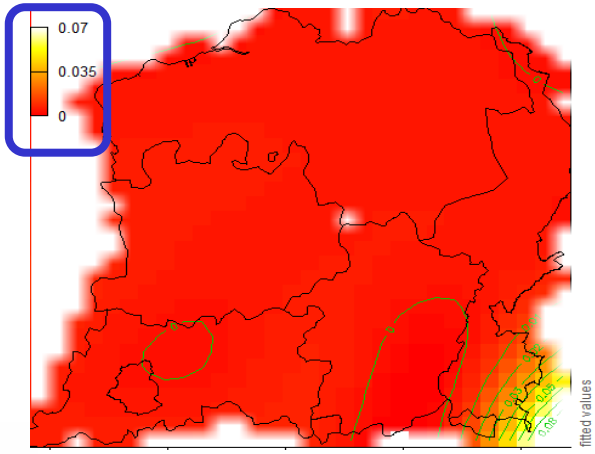
clothing



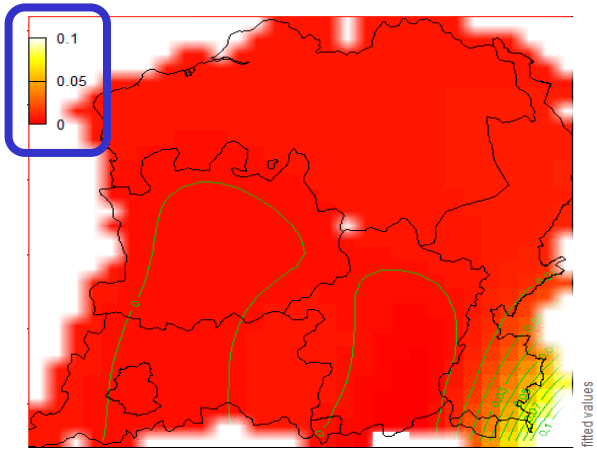
personality & feelings



church & religion



society, school & education

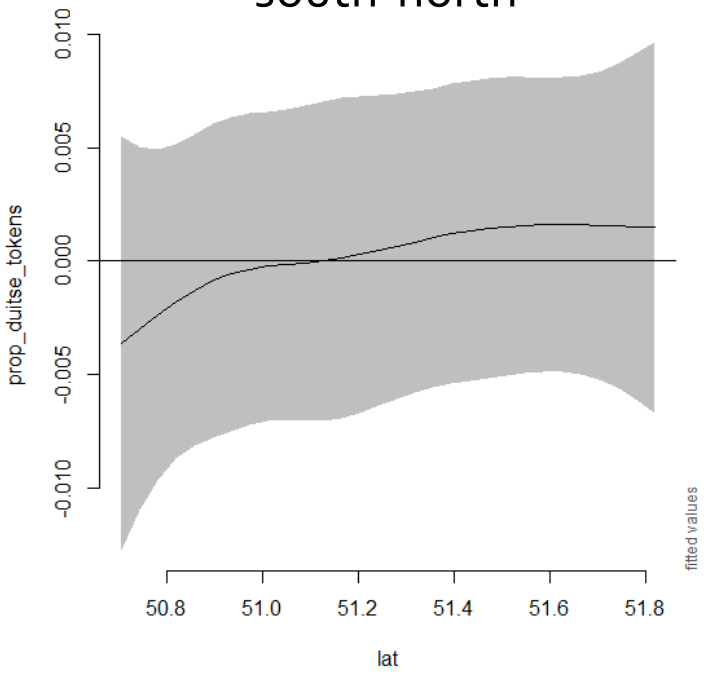


deviance explained : 90.4%

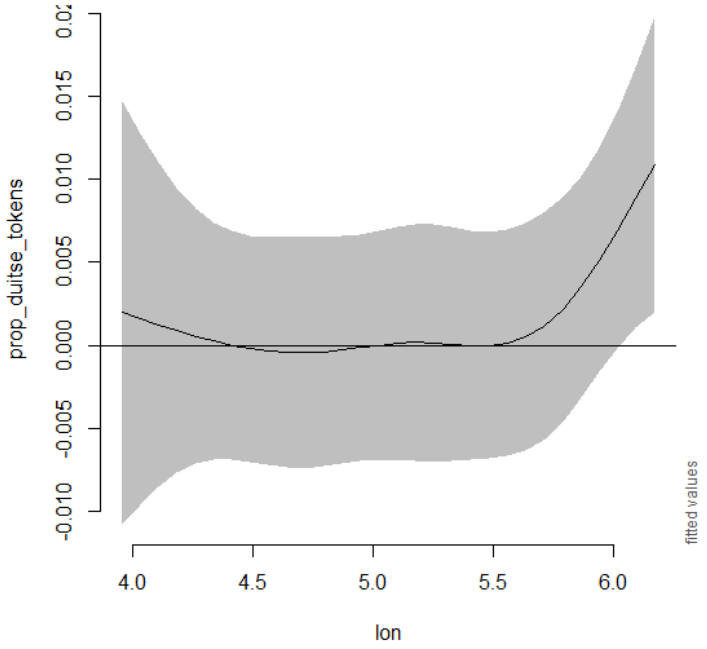
geographical patterns: German

model struggles with general infrequency of German

south-north



west-east



discussion

- expectations partly confirmed:
  - more French in Flanders especially for clothing terminology
  - geography affects the use of Latin, but semantics is more important for borrowings from this source language
  - more German near the German border, but German is only frequently used in a few locations

- cultural-historical background reflected in variation in naming
- naming strategies also affect the amount of geographical heterogeneity in dialects

  e.g. homogeneity for concepts relating to church & religion

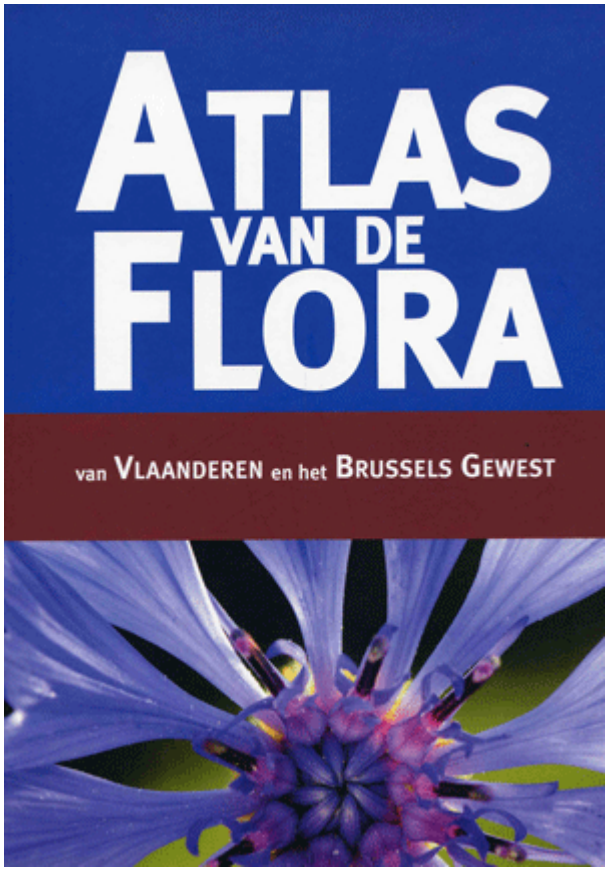
4.
let's talk about plants, baby

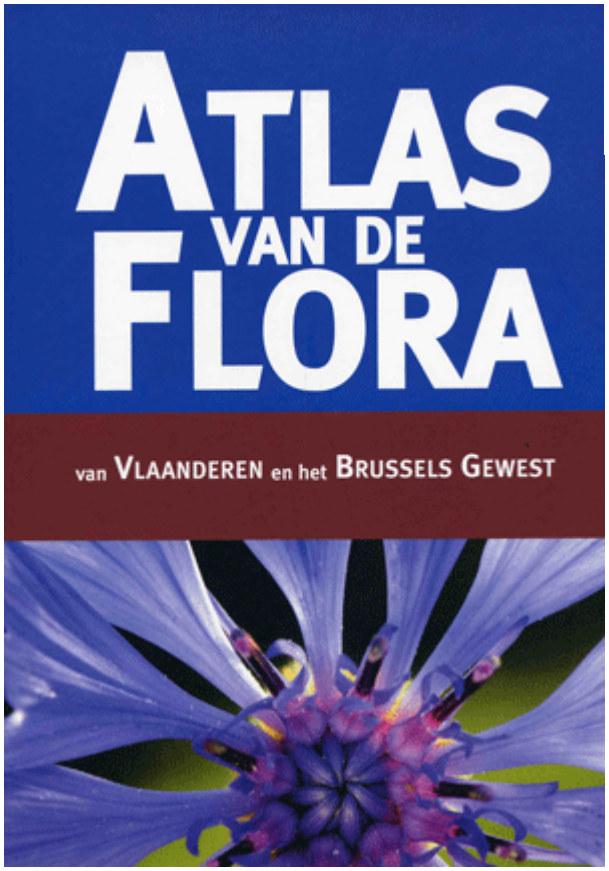
correlating experiential salience and lexical variation

Experiential salience

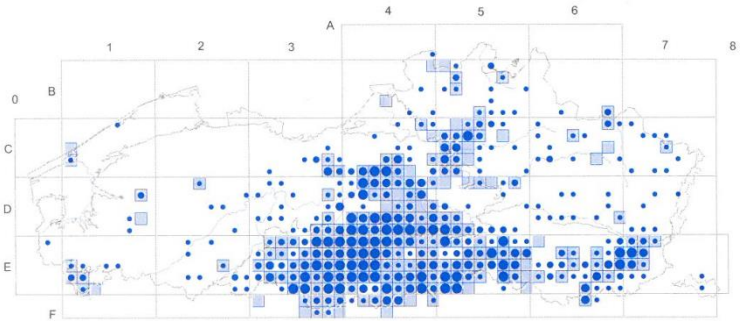
1. referential frequency of a concept
2. extension: folkloristic relevance of a concept

→ investigating plant name variation

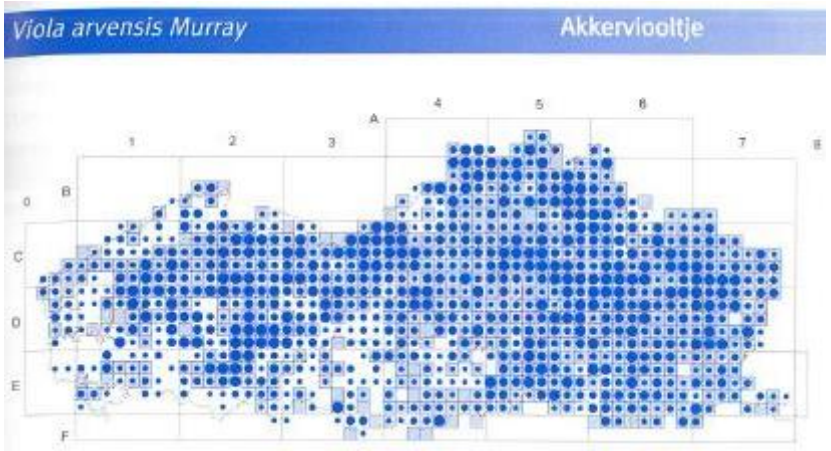

ATLAS
VAN DE
FLORA
van VLAANDEREN en het BRUSSELS GEWEST



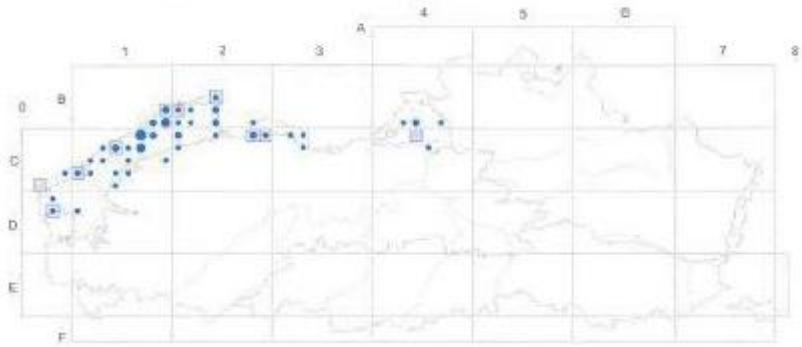
Cirsium oleraceum (L.) Scop. **Moesdistel**



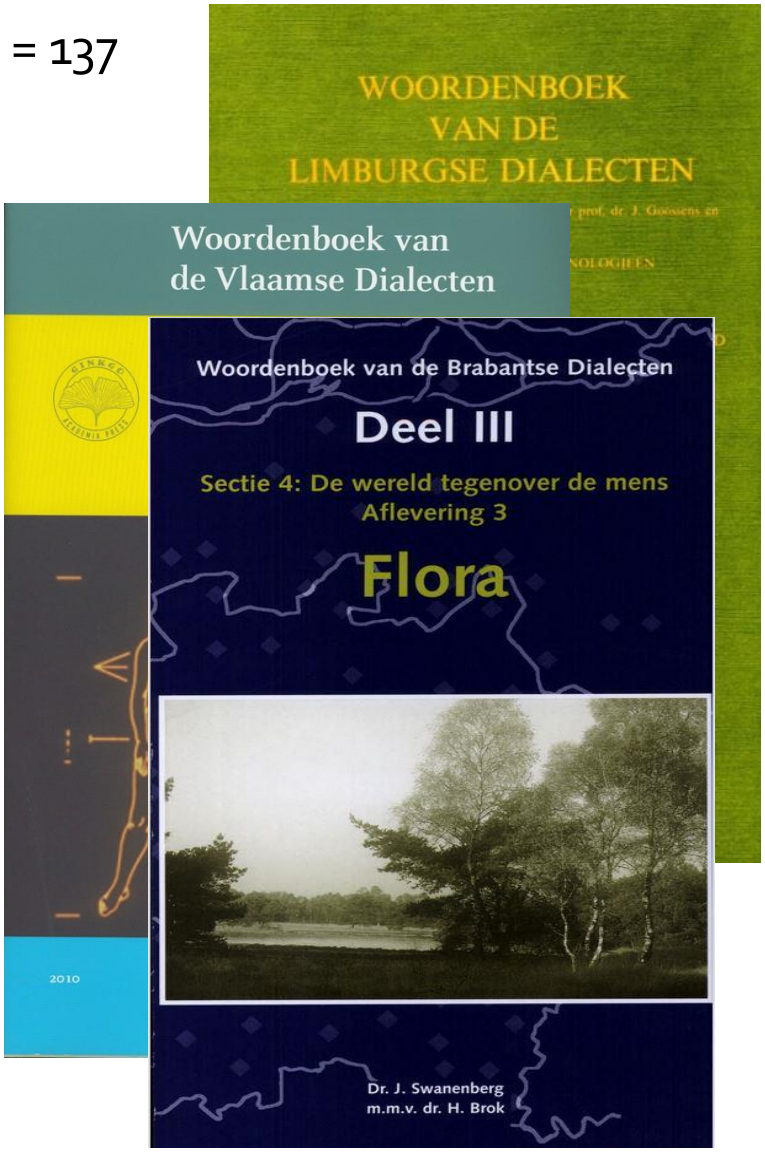
Viola arvensis Murray **Akkervlooltje**



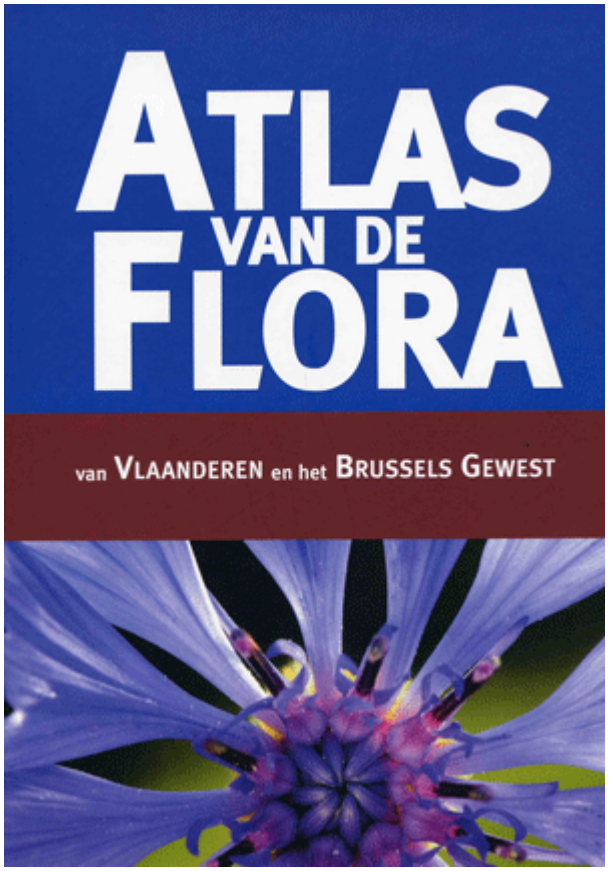
Ranunculus baudotii Godr. **Zilte waterranonkel**



N = 137


WOORDENBOEK
VAN DE
LIMBURGSE DIALECTEN
Woordenboek van
de Vlaamse Dialecten
2010
Woordenboek van de Brabantse Dialecten
Deel III
Sectie 4: De wereld tegenover de mens
Aflevering 3
Flora
Dr. J. Swanenberg
m.m.v. dr. H. Brok

N = 137

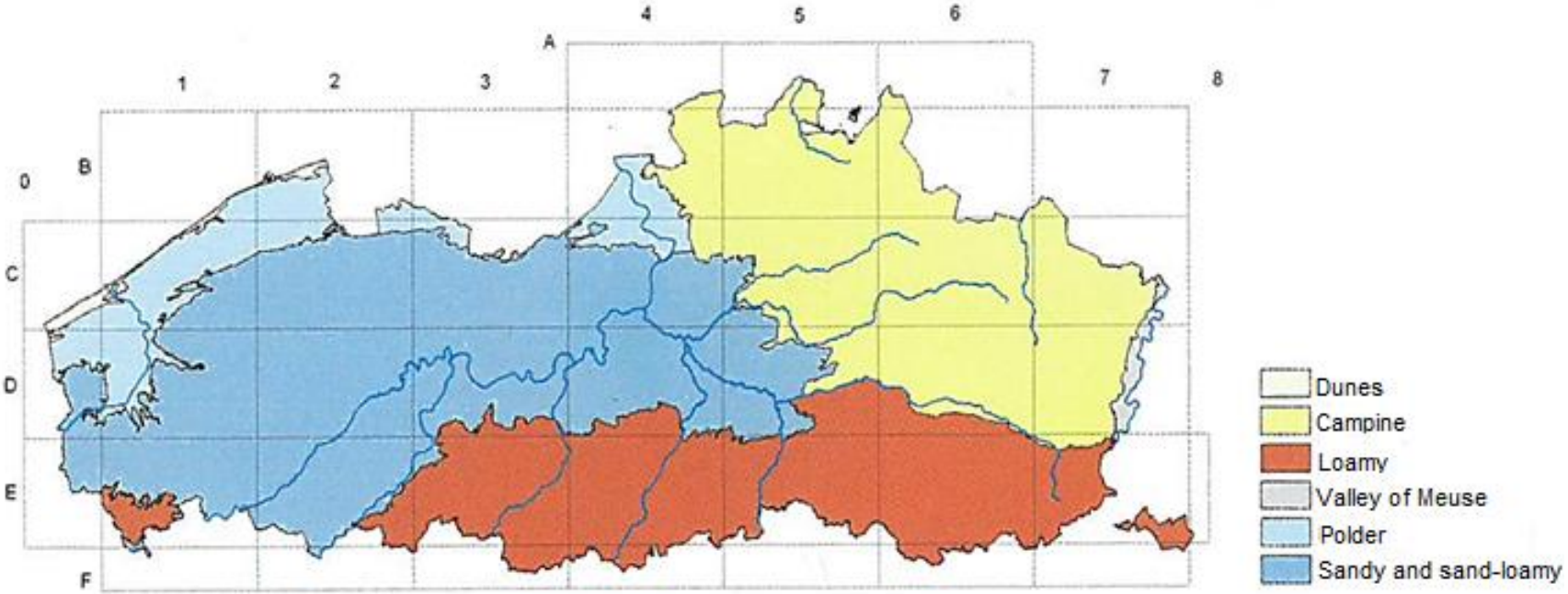

ATLAS
VAN DE
FLORA
van VLAANDEREN en het BRUSSELS GEWEST

?


WOORDENBOEK
VAN DE
LIMBURGSE DIALECTEN
Woordenboek van
de Vlaamse Dialecten
Woordenboek van de Brabantse Dialecten
Deel III
Sectie 4: De wereld tegenover de mens
Aflevering 3
Flora
2010
Dr. J. Swanenberg
m.m.v. dr. H. Brok

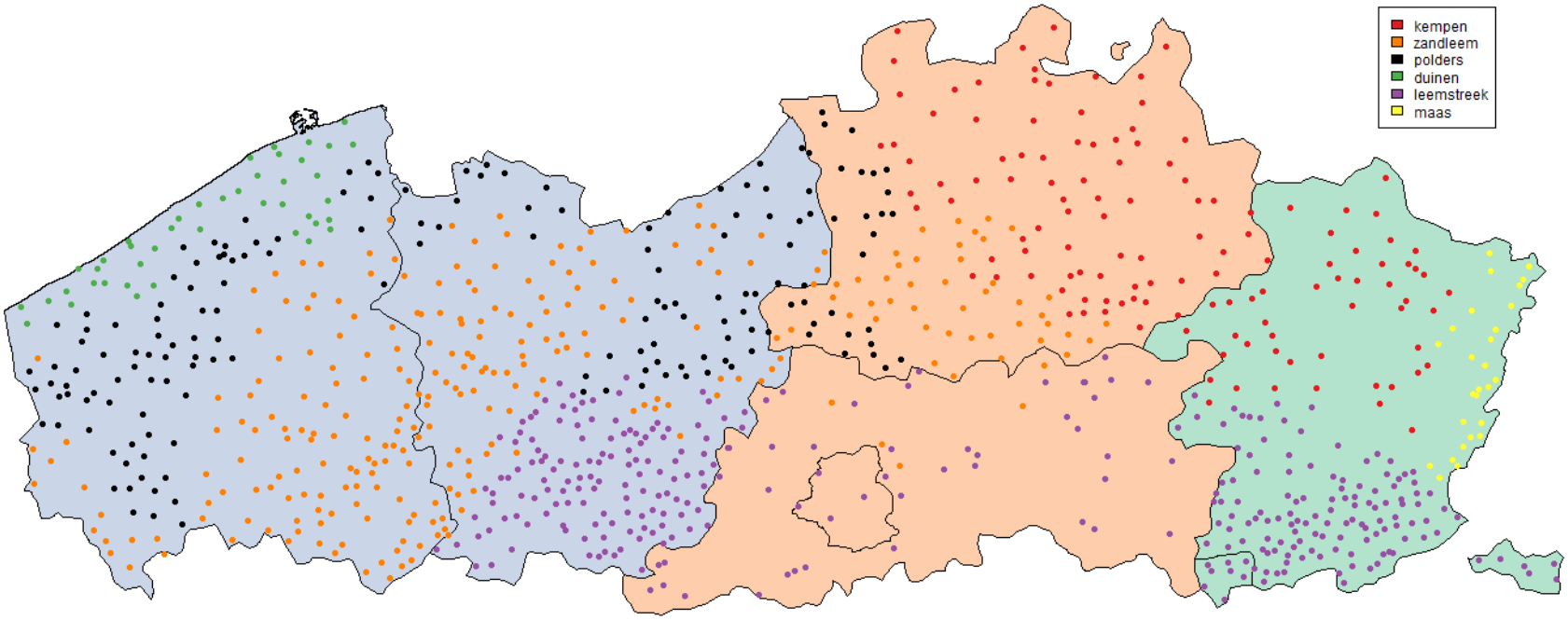
calculating lexical diversity

– calculated per plant per ecological region



calculating lexical diversity

– calculated per plant per ecological region



WVD WBD WLD

calculating lexical diversity

- calculated per plant per ecological region

- type-token ratio (TTR):
  - number of different lexemes (types) / number of records (tokens)
  - higher value = more variation
  - 30% of data: number of types = number of tokens (max = 11)

- internal uniformity (I; Geeraerts, Grondelaers & Speelman 1999):
  - $I_Z(Y) = \sum_{i=1}^{n} F_{Z,Y}(x_i)^2$
  - takes into account frequency of different lexemes and relative frequency of each lexeme
  - lower value = more variation

internal uniformity (I)

vergeet-mij-niet(je):
93.55% (N = 232)

blauwe kanne:
0.8% (N = 2)

onzevrouwetraantjes:
0.8% (N = 2)

…
(8 lexemes with N = 2)

$I = 0.9355^2 + 8 * (0.008^2)$
$= 0.8757$

internal uniformity (I)

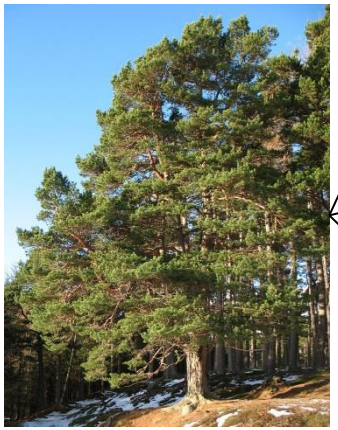
vergeet-mij-niet(je):
93.55% (N = 232)

blauwe kanne:
0.8% (N = 2)

onzevrouwetraantjes:
0.8% (N = 2)

…
(8 lexemes with N = 2)

$I = 0.9355^2 + 8 * (0.008^2)$
$= 0.8757$



den:
62.5% (N = 10)

grove den:
6.25% (N = 1)

mast:
31.25% (N = 5)

$I = 0.625^2 + 0.0625^2 + 0.3125^2$
$= 0.4922$

combining the referential and linguistic data

calculated per plant per ecological region:

plant	ecological region	global frequency 1 (abs. freq.)	global frequency 2 (abs. freq.)	global frequency 3 (abs. freq.)	local frequency (rel. freq.)	number of records	different lexemes	TTR	I
beech	Campine	2229	248	678	25.2	4	2	0.500	0.500
beech	Dunes	2229	248	678	14.6	24	3	0.125	0.462
beech	Loamy	2229	248	678	46.5	97	5	0.052	0.758
beech	Polder	2229	248	678	1.9	175	5	0.029	0.574
beech	Sand-loamy	2229	248	678	25.1	433	9	0.021	0.616

combining the referential and linguistic data

calculated per plant per ecological region:

plant	ecological region	global frequency 1 (abs. freq.)	global frequency 2 (abs. freq.)	global frequency 3 (abs. freq.)	local frequency (rel. freq.)	number of records	different lexemes	TTR	I
beech	Campine	2229	248	678	25.2	4	2	0.500	0.500
beech	Dunes	2229	248	678	14.6	24	3	0.125	0.462
beech	Loamy	2229	248	678	46.5	97	5	0.052	0.758
beech	Polder	2229	248	678	1.9	175	5	0.029	0.574
beech	Sand-loamy	2229	248	678	25.1	433	9	0.021	0.616

methods & expectation

negative correlation plant frequency & lexical variation:

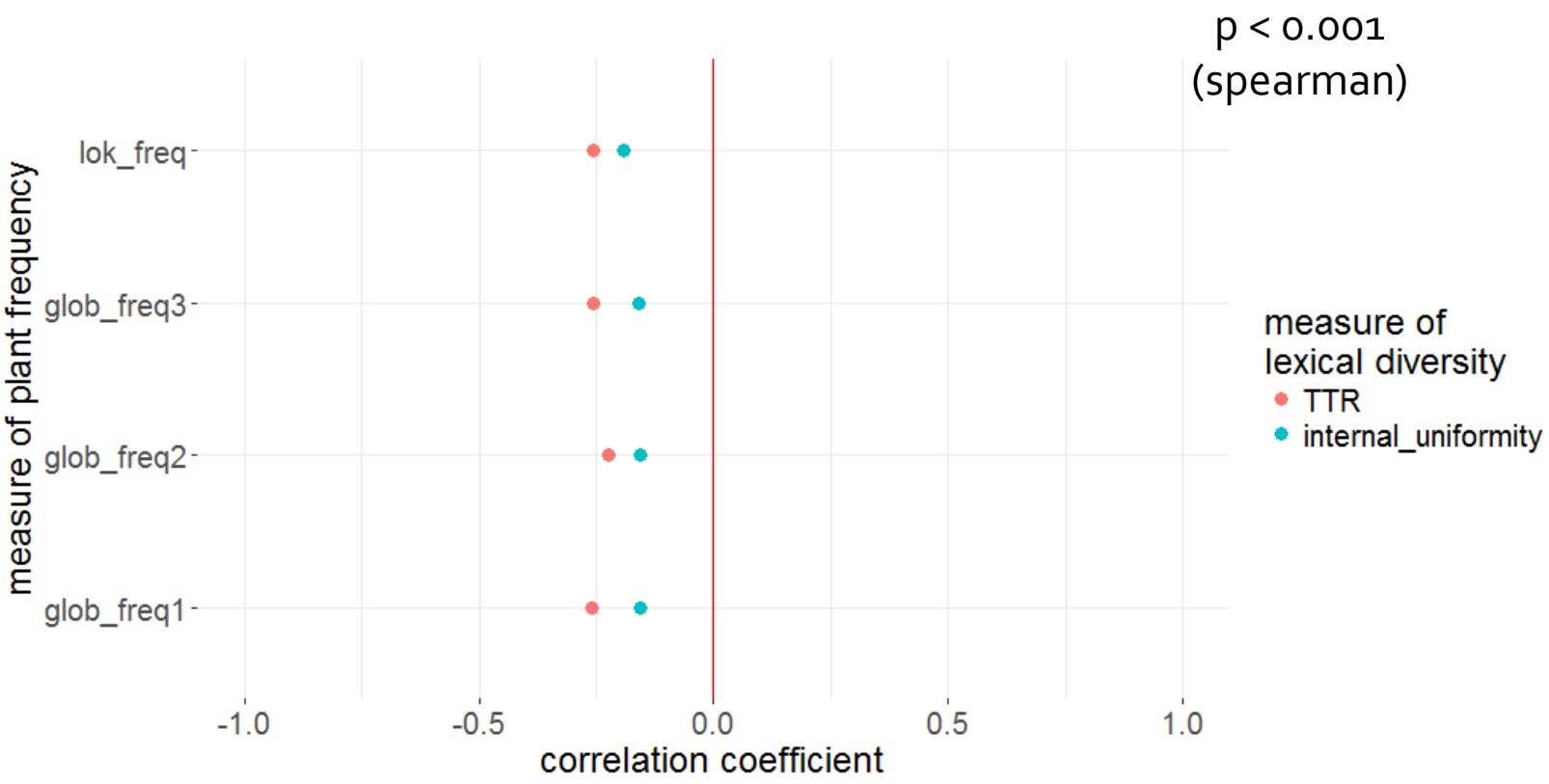
- spearman rank correlation tests
- correlation coefficients

→ TTR: negative correlations expected

internal uniformity: positive correlations expected

results

frequency measures * lexical variation



discussion

- TTR: results as expected

  significant negative correlation between plant frequency & lexical variation

  → less frequent plants show more lexical variation

- internal uniformity: results show opposite effect

  → names for frequent plants are not standardized enough to be picked up by I

why these diverging results?

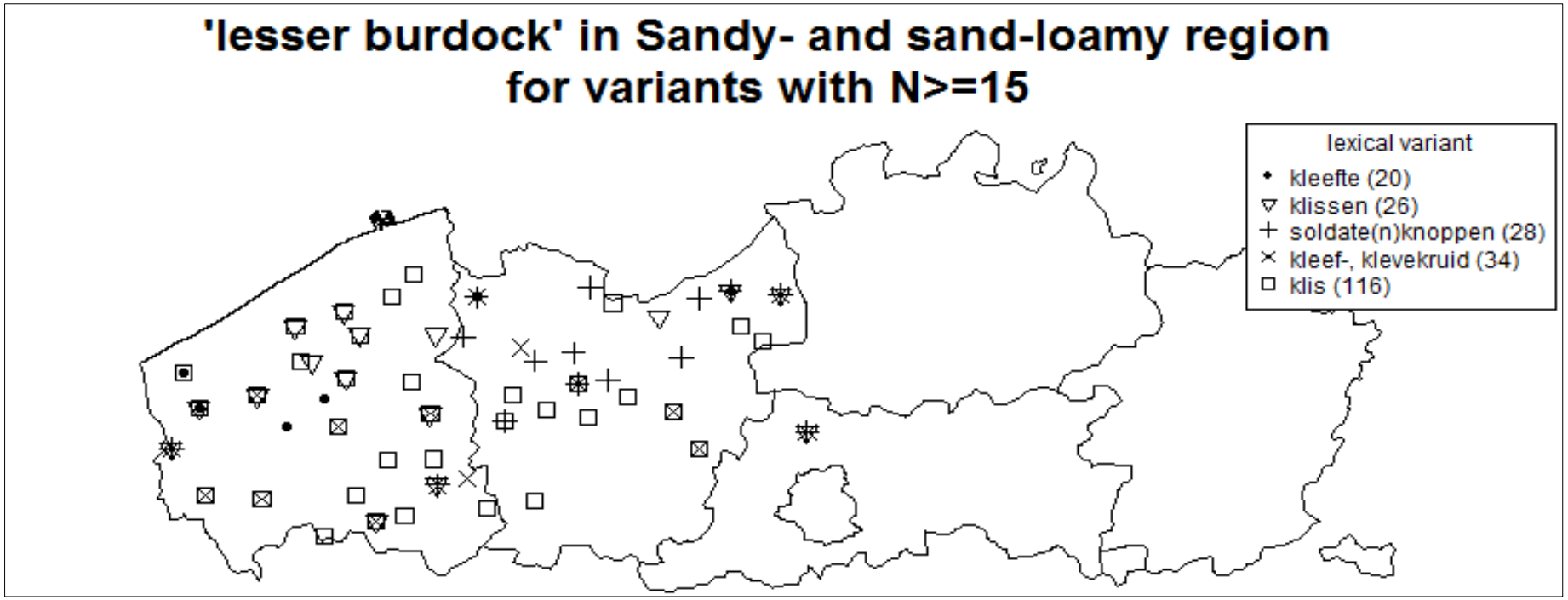
1. TTR and internal uniformity measure conceptually different phenomena
2. ecological regions vs. dialect regions

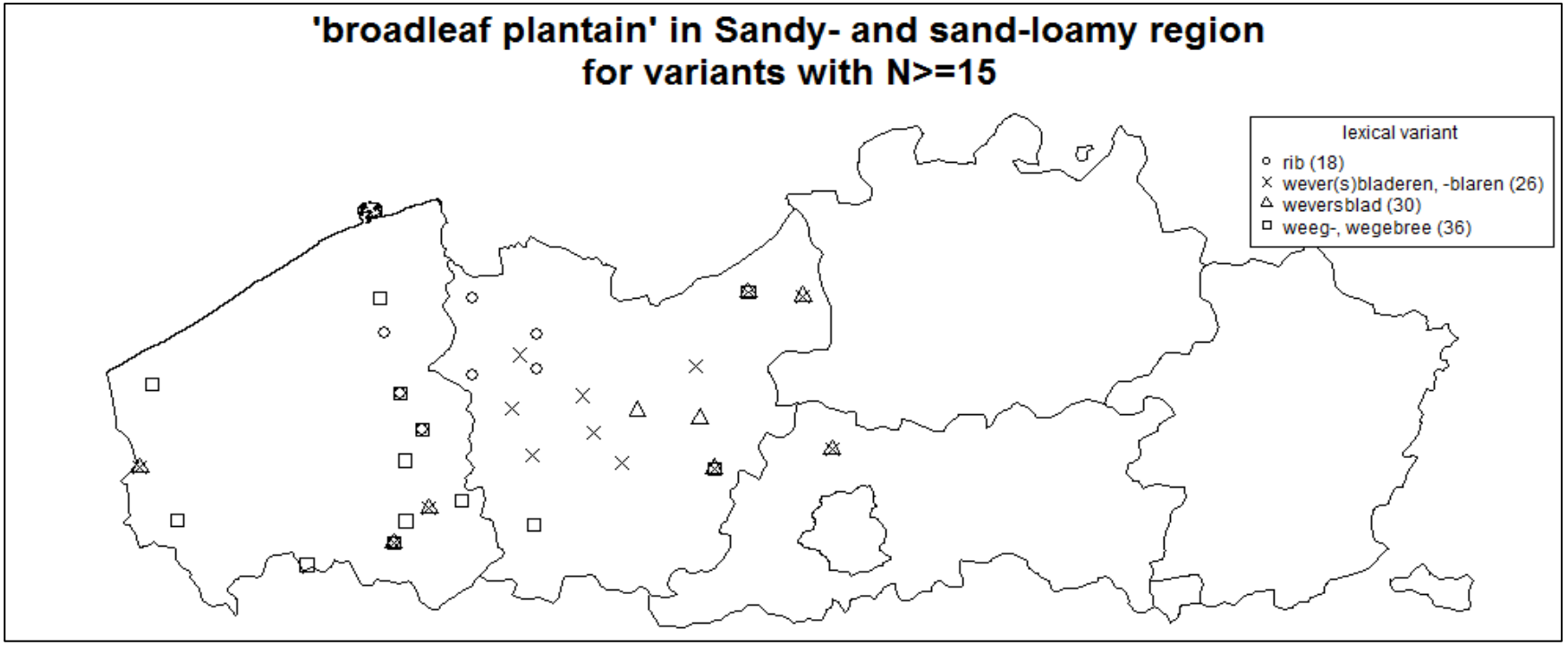
TTR vs. I

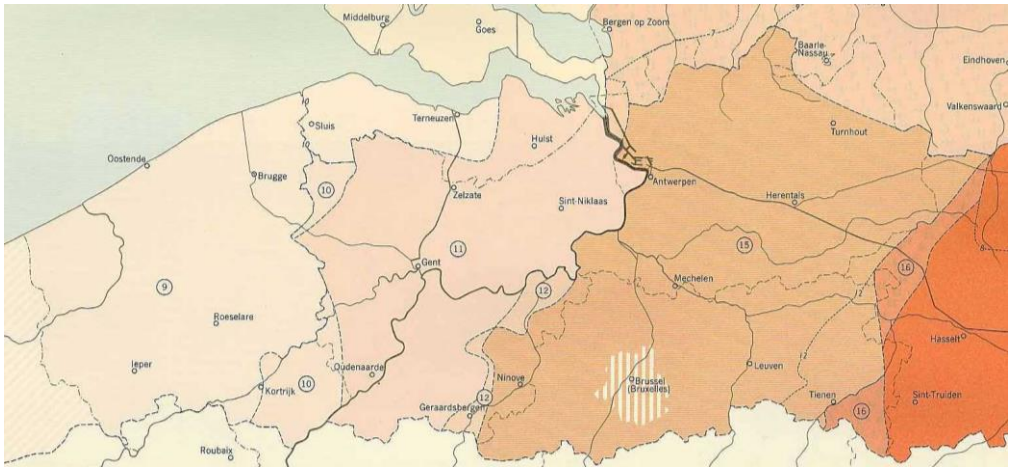
plant (ecological region)	number of records	distribution of types	nr. of diff. lexemes	TTR	I
great mullein, Loamy region	26	$lexeme_{1...18}$ occur once $lexeme_{19...22}$ occur once	22	0.846	0.050
bitter dock, Polder region	38	$lexeme_{1,2}$ occur once $lexeme_{3}$ occurs 3 times $lexeme_{4}$ occurs 4 times $lexeme_{5}$ occurs 10 times $lexeme_{6}$ occurs 19 times	6	0.158	0.338
black locust, Sandy and sand-loamy region	26	$lexeme_{1,2,3}$ occur once $lexeme_{4}$ occurs 23 times	4	0.154	0.787
forget-me-not, Dunes region	52	$lexeme_{1}$ occurs 52 times	1	0.019	1

TTR vs. I

plant (ecological region)	number of records	distribution of types	nr. of diff. lexemes	TTR	I
great mullein, Loamy region	26	$lexeme_{1...18}$ occur once $lexeme_{19...22}$ occur once	22	0.846	0.050
bitter dock, Polder region	38	$lexeme_{1,2}$ occur once $lexeme_{3}$ occurs 3 times $lexeme_{4}$ occurs 4 times $lexeme_{5}$ occurs 10 times $lexeme_{6}$ occurs 19 times	6	0.158	0.338
black locust, Sandy and sand-loamy region	26	$lexeme_{1,2,3}$ occur once $lexeme_{4}$ occurs 23 times	4	0.154	0.787
forget-me-not, Dunes region	52	$lexeme_{1}$ occurs 52 times	1	0.019	1


'lesser burdock' in Sandy- and sand-loamy region
for variants with N>=15
lexical variant
kleefte (20)
klissen (26)
soldate(n)knoppen (28)
kleef-, klevekruid (34)
klis (116)





Daan & Blok 1969

9: West-Flemish & Zeelandic Flemish

10: intermediate dialects between West- and East-Flemish

11: East-Flemish

15: Brabantic

further research

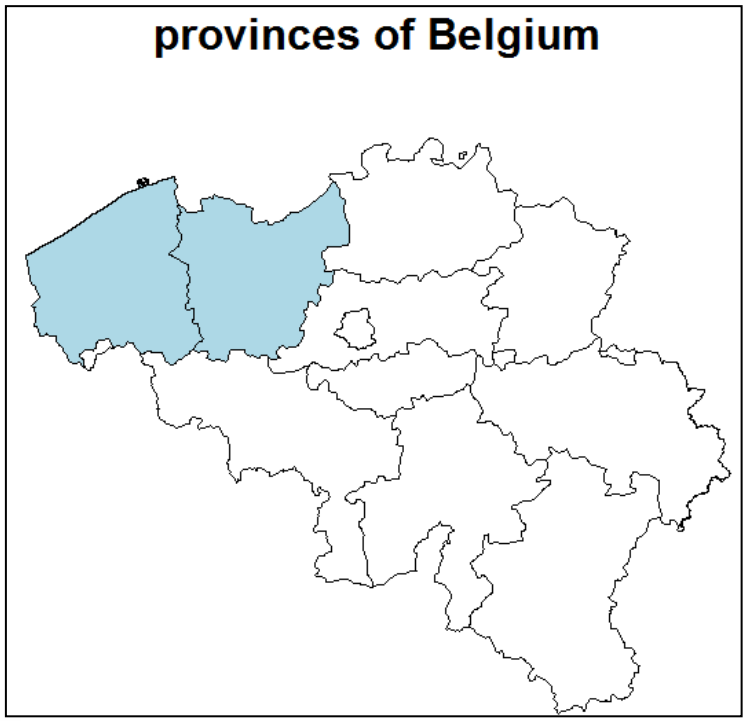
- restrictions on the data set

- small effect sizes
  - all plants relatively frequent
  - data from other language areas? other measures of experiential salience?

data from other language areas

combining dialect dictionaries from two languages

- dictionary of the Flemish dialects (WVD: dialects of Dutch in west of Flanders)
- DBÖ (Bavarian Dialects of Austria)



other measures of experiential salience

- referential plant frequency (Atlas & GBIF)
- edibility rating (pfaf.org)
- medicinal rating (pfaf.org)
- poisonousness (data U Cornell)

- hypothesis: the more experientially salient the plant, the smaller the amount of lexical variation

  → less variation for plants that…

  are more frequent

  have a higher edibility rating

  have a higher medicinal rating

  are poisonous (vs. not poisonous)

results (TTR)

- referentially more frequent plants show a significantly smaller amount of lexical variation (spearman $p < 0.01$, $r = -0.310$)

  → opposite effect in Bavarian data

- edible plants show a significantly smaller amount of lexical variation ($p < 0.01$, Adj R^2: 0.065)

  → similar trend in Bavarian data (NS)

- plants that are useful for medicinal applications show a significantly smaller amount of lexical variation ($p < 0.05$, Adj R^2: 0.039)

  → similar trend in Bavarian data (NS)

- the poisonousness of a plant does not have any significant effect, but on average, poisonous plants show more variation

discussion

- experiential salience influences the amount of lexical variation in dialect data
  - referential frequency
  - folkloristic relevance
- further research: correlation with text-based frequency

  → what makes a concept salient?

conclusions (part 1)

the effect of cognitive concept features
on lexical geographical variation is stable

- it persists in other semantic fields than the human body
- it cannot solely be explained by the geographical signal in the data

- semantic fields
  - can be arranged along an axis of degree of universality:

    local > society-related > universal
  - some fields are more prone to geographical fragmentation than others

conclusions (part 2)

social and cultural features also affect the structure of lexical dialect variation

- the socio-historical background of a language user interacts with lexical geographical variation
- naming strategies reflect semantic and geographical structure
- experiential salience correlates with lexical variation

what does this mean?

- for (lexical) dialectometry and for studies in lexical variation in other types of stratificational varieties:
  - dialectometric results will be influenced by concept-related features (see Speelman & Geeraerts 2008)
  - traditional dialectologists are probably (implicitly) aware of these features, but they are rarely ever explicitly accounted for

- for Cognitive (Socio-)linguistics:
  - language variation (and change?) is clearly affected by features that are related to the mental organization of the lexicon (part 1)
  - these features are influenced by the everyday environment and socio-historical background of a language user (part2)

Thank you!

Questions? Suggestions?

extra

response case-studies 1 & 2

lexical diversity

calculated as the number of types per concept

e.g. TO GET MARRIED (TROUWEN): 3 different types

trouwen	181
zich binden	1
getrouwd worden	1

WELL-BUILT WOMAN (GROF GEBOUWDE VROUW): 131 different types

machochel	67	mokkel	8
schommel	41	bai (fr.)	7
molenpaard	23	madsel	5
machine	17	schokkel	5
kapitein	11	dikke madam	4
mangel	11		…

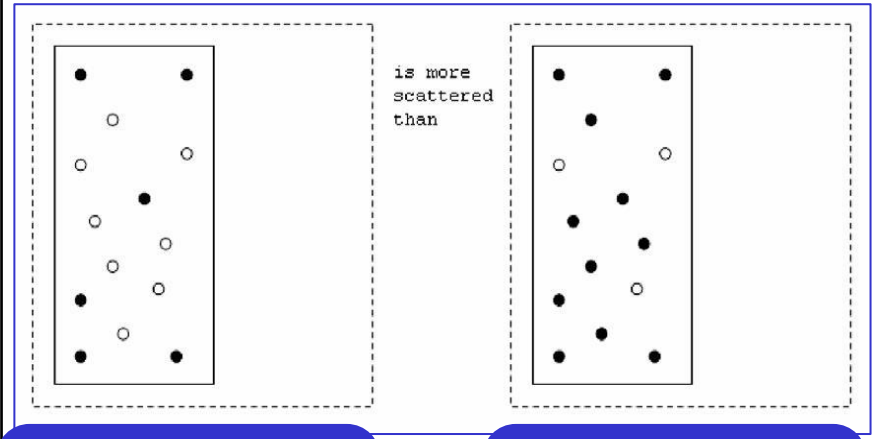
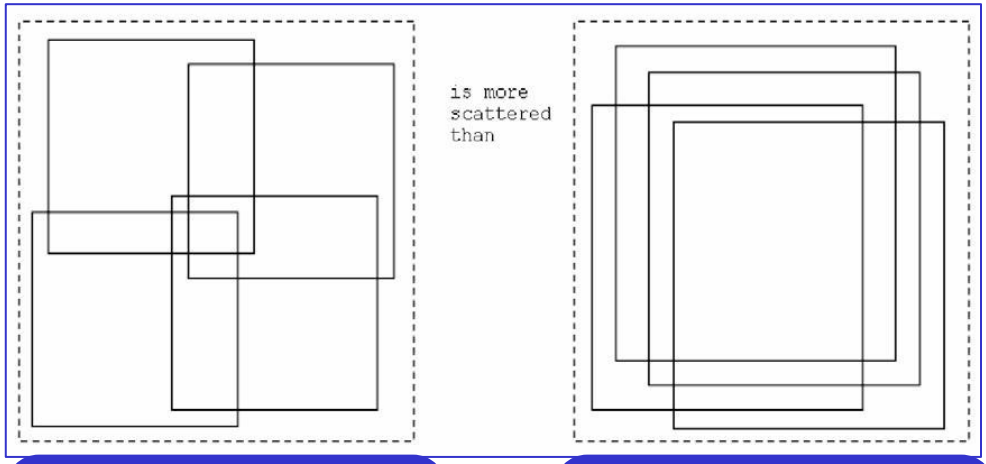
geographical fragmentation

- calculated as the proportion of dispersion and range

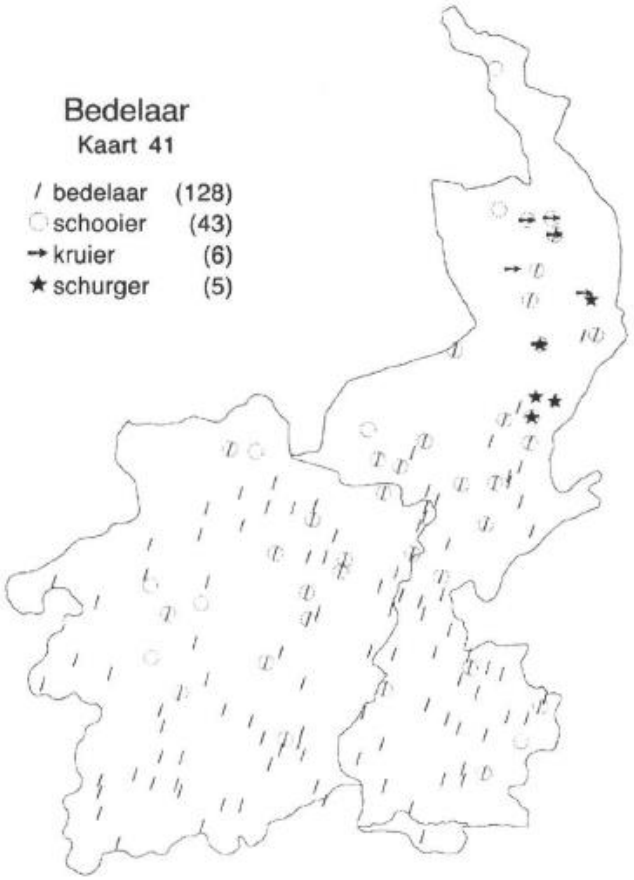
- dispersion: (weighted) average distance between the attestations of the unique words for a concept relative to other words for the same concept
- range: (weighted) average coverage of the words for a concept relative to the entire region where the concept occurs

(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)

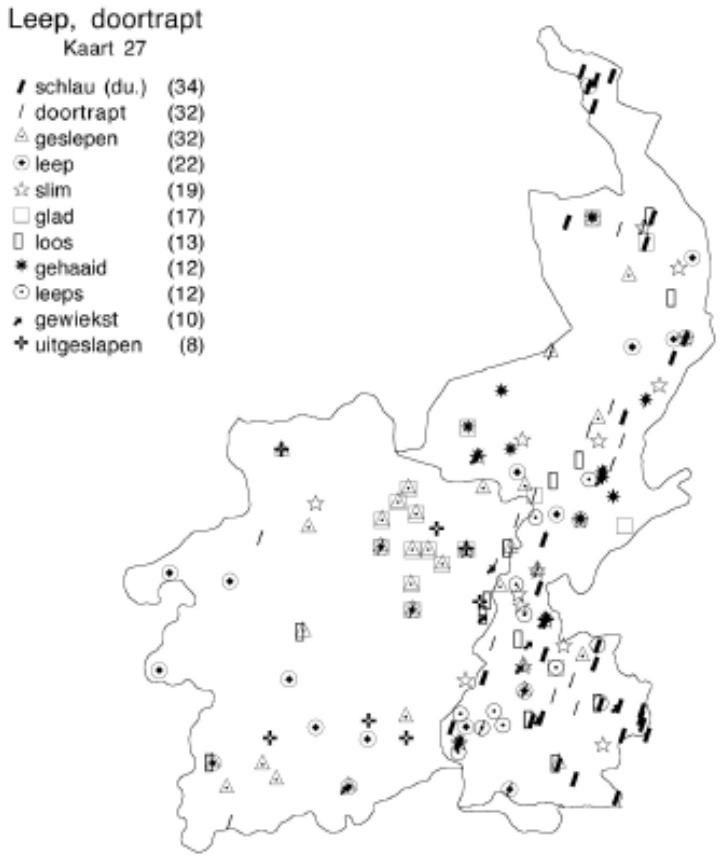
dispersion & range

dispersion	range
	
variants scattered across dialect area variants are found in nearby locations	each word type occurs in small geographical area each word type takes up almost entire dialect area

dispersion

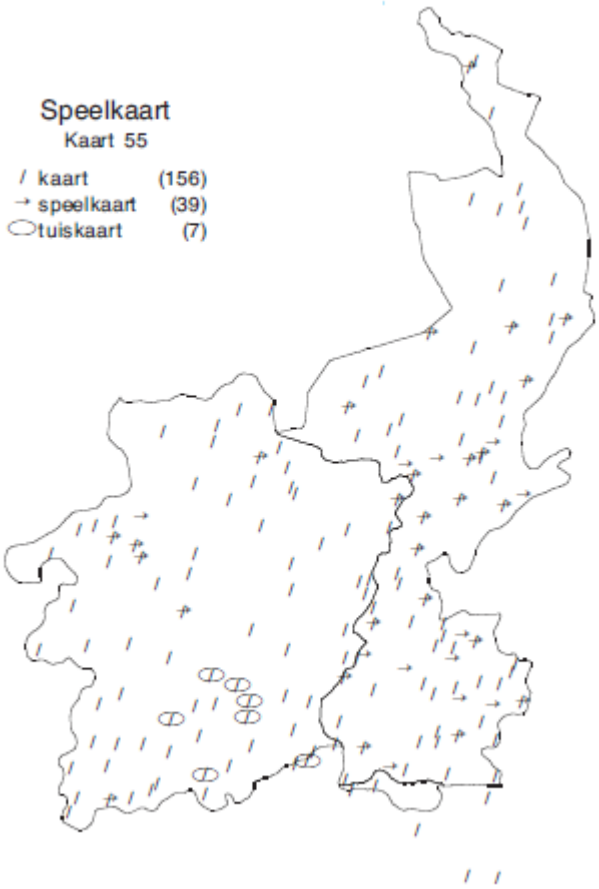


dispersion = 1.22

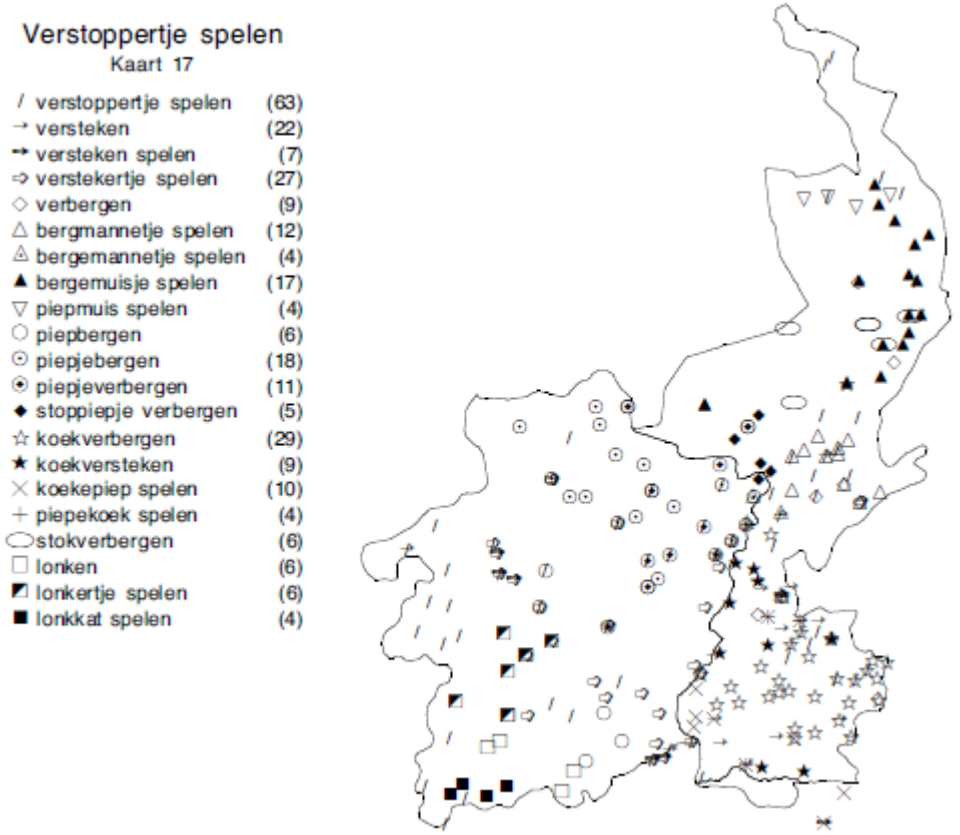


dispersion = 2.58

range



range = 0.82



range = 0.20

predictors case-studies 1 & 2

concept-related predictors

1. LACK OF SALIENCE
   - proportion of missing places
     → ambiguous
   - proportion of multi-word expressions (MWE)
   - proportion of hapax legomena
   - prevalence (Keuleers et al. 2015)
     - word-level
     - missing data
2. VAGUENESS
   - number of types also used for other concepts (GS10, SG08)
3. AFFECT
   - manual, but relatively stable
   - mean valence (Moors et al. 2013), but missing data